DOCUMENT RESUME

ED 223 682 TM 820 819

AUTHOR	Holmes, Mark
TITLE	What Every Teacher and Parent Should Know about Student Evaluation. Informal Series No. 46.
INSTITUTION	Ontario Inst. for Studies in Education, Toronto. Dept. of Educational Planning.
REPORT NO	ISBN-0-7744-5058-4
PUB DATE	82
NOTE	147p.
PUB TYPE	Guides - Non-Classroom Use (055)
EDRS PRICE	MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS	Administrator Responsibility; *Educational Policy; Elementary Secondary Education; Evaluation Criteria; *Evaluation Methods; Foreign Countries; Measures (Individuals); School Community Relationship; *Student Evaluation; Teacher Made Tests; Test Construction; *Testing; Testing Problems
IDENTIFIERS	Ontario

ABSTRACT

This handbook is intended primarily for the classroom teacher and the school principal. The focus of the book is on the evaluation of students as they pass through elementary and secondary school. It includes suggestions for the classroom teacher in developing tests and other assessment measures. It deals with a range of educational objectives in the intellectual, cultural, aesthetic, biological, and moral domains. It also examines problems of policy from the point of view of the school staff as a whole, the principal, and superintendent. The contents include: (1) why evaluate, (2) tests, (3) improving skills in evaluation, (4) planning and organizing student evaluation, and (5) the school and the public. (Author/PN)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

What Every Teacher and Parent Should Know about Student Evaluation

Mark Holmes
Assistant Director, Planning and Resources
The Ontario Institute for Studies in Education

Informal Series/46

OISE Press/The Ontario Institute for Studies in Education

The Ontario Institute for Studies in Education has three prime functions: to conduct programs of graduate study in education, to undertake research in education, and to assist in the implementation of the findings of educational studies. The Institute is a college chartered by an Act of the Ontario Legislature in 1965. It is affiliated with the University of Toronto for graduate studies purposes.

The publications program of the Institute has been established to make available information and materials arising from studies in education, to foster the spirit of critical inquiry, and to provide a forum for the exchange of ideas about education. The opinions expressed should be viewed as those of the contributors.

©The Ontario Institute for Studies in Education 1982
252 Bloor Street West
Toronto, Ontario
M5S 1V6

All rights reserved. No part of this publication may be reproduced in any form without permission from the publisher, except for brief passages quoted for review purposes.

Canadian Cataloguing in Publication Data
Holmes, Mark, 1935-
What every teacher and parent should know about student evaluation

(Informal series ; 46)
Co-published by Ontario Institute for Studies in Education.
ISBN 0-7744-5058-4

1. Students, Rating of. 2. Educational tests and measurements. I. Ontario Institute for Studies in Education. II. Title. III. Series: Informal series (Ontario Institute for Studies in Education) ; 46.

LB 3051-H64 371.2'7 C82-094886-1

ISBN 0-7744-5058-4 Printed in Canada
1 2 3 4 5 XX 68 58 48 38 28

TABLE OF CONTENTS

Introduction/1

1 Why Evaluate?/3

2 Tests/17

3 Improving Skills in Evaluation/40

4 Planning and Organizing Student Evaluation/99

5 The School and the Public/135

Index/139

INTRODUCTION

This handbook is intended primarily for the classroom teacher and the school principal. It assumes no particular statistical competence or interest. I suspect that many people avoid the problem of evaluation because they wrongly believe that only mathematicians can understand it. I hope also that concerned parents will find this handbook useful. As parents become more informed, they will be able to understand better some of the daily problems teachers face and they will be able to frame questions and doubts more intelligently.

This book makes no attempt to be objective about different approaches to evaluation. I do not maintain the pretence that any position on any topic is as sensible as any other. I do hold the view that there is good and bad practice in evaluation, as there is in education generally, and this book is intended to improve the general standard. Nevertheless, I have tried to allow, as far as reasonably possible, room for discussion of different points of view on the more controversial questions, at the same time making my own recommendations clear.

The focus of the book is on the evaluation of students as they pass through elementary and secondary school. It includes suggestions for the classroom teacher in developing tests and other assessment measures. It deals with a range of educational objectives in the intellectual, cultural, aesthetic, biological, and moral domains. It also examines problems of policy from the point of view of the school staff as a whole, the principal and superintendent. I have tried to deal with both elementary and secondary problems but the examples come more from the middle grades than from either extreme.

Most of the ideas and recommendations stem from my own experience, as a teacher and administrator in New Brunswick, as an administrator in Quebec, and as a professor of education heavily involved in the field in Ontario. I am particularly grateful to the teachers and administrators with whom I have worked in all three provinces for helping me think through many of the problems. The blame for all the book's inadequacies is, of course, mine.

1 WHY EVALUATE?

The Purposes of Evaluation in the School Classroom

This chapter is intended to explain just why evaluation is so important in the classroom. Not only do different teachers evaluate for different purposes, but the same teachers evaluate for various purposes. As this book is primarily concerned with the evaluation of students, individually and in groups, the purposes considered here are those the individual teacher is likely to have. Consideration is not given to evaluation for purposes of national or provincial assessment.

The purposes of the classroom teacher can be categorized under four headings--assessment for instructional purposes; assessment as a means of providing incentives for learners; assessment as a means of classifying learners; and assessment for purposes of certification.

Assessment for Instructional Purposes

Of all the reasons for evaluating learners, this is the least controversial. Almost every teacher watches facial expressions, asks questions, and gives short tests to gauge just what class members have learned and what they have not learned. Increasingly, too, research supports the traditional adage, "Teach, test, review." Teachers frequently begin a unit by finding out, formally or informally, just what the pupils already know and understand. They build on the strengths and try to eradicate weaknesses.

One of the indicators of poor teaching is a tendency to drift from topic to topic, from unit to unit, without ever ascertaining precisely what has and what has not been learned. In those circumstances, the latter category is likely to be considerably fuller than the former.

Consider this example. Suppose one is teaching the geography of the Atlantic Provinces to one of the middle grades. There follow first a bad example and then a good example of the use of instructional assessment.

Bad Example:

Lesson 1 - The teacher talks about the Atlantic Provinces and asks a few questions. The teacher uses a wall map to show the provinces and the major cities and rivers. One of the pupils describes a vacation trip made to Nova Scotia.

Lesson 2 - A film is shown on the Atlantic Provinces.

Lesson 3 - The teacher talks about fishing, mining, and the pulp and paper industry in the Atlantic Provinces, and tells the class it is to have a test on what "they have covered so far" the following day. When asked what the pupils are "responsible for," the teacher replies, "Everything we've done so far."

Lesson 4 - The next day's test consists of mainly factual questions. Some answers have been given by the teacher during the classes. Most of the remainder can be found in the text book. A few are tricky questions of minor details "to catch out those who are not paying attention." Most of the pupils do very badly on the test, the exceptions being those who memorized the chapter in the text.

The tests are returned to the pupils two weeks later with sarcastic remarks about the pupils' laziness. Another test is given later with most of the same questions. The pupils have memorized the answers from the first test and perform much better on the second. The two tests are averaged with a homework assignment to provide a mark for the first term report.

What's wrong with that?

1. The teacher never explained the objectives of the unit.
2. The teacher never found out what the pupils already knew.
3. The teacher never explained the relevance of the film.

4. The teacher never explained what the test was on or what its purpose was.
5. The only successful students were those who had memorized the largest number of facts without being told to do so. In other words, they knew the system and it had little to do with the stated objectives.
6. There was no immediate feedback from the test.
7. Even from the second test, little was determined beyond the proportion of specific answers that had been memorized.

Good Example:

Lesson 1 - The teacher begins by briefly explaining the objectives of the unit. A quick test is given measuring knowledge of a few basic facts (names of provinces, lines of latitude, major cities, major products, climate) and one or two geographical concepts (e.g. relationship between proximity to the ocean and temperature in summer and winter; the relationship between topographical features and transport patterns; the relationship between climate and agriculture). The teacher has the tests marked immediately, assuring pupils that it will not count. The teacher shows no displeasure at the discovery of almost total ignorance, commenting only that there is plenty of work for him to do. "Now we know the gaps," he says cheerfully, "let's see how well we can fill them."
The teacher then hands the pupils a written statement of objectives and explains:

1. What each means;
2. How each objective can be achieved;
3. The dates by which the objectives should be achieved.

There is general discussion of what kind of things will have to be known or understood for basic competence (passing), and what kind of things for excellence (an 'A').

Lesson 2 - Using a map of the world, one of Canada, and one of the Atlantic Provinces, the teacher shows the geographical situation of the Atlantic Provinces, their rivers and major centres, and their links with other parts of Canada, and with the world. There is a general discussion about place, resources, lines of communication, and the things that make Atlantic centres different from various other parts of Canada already studied.

Lessons 3, 4, 5, 6 - The teacher teaches the various concepts, as applied to the Atlantic Provinces. Each is checked with a follow-up quiz and the learning is reinforced with a quick review the following day. Although the class has to "discover" certain relationships, it is clear what *must* be remembered, what *should* be remembered, and what would be *nice* to know. Maps are provided or drawn with the important facts.

Lesson 7 - In introducing the film, The Atlantic Provinces, the teacher tells the class there will be a quiz after the film. He reminds them that geography is about the difference between places. He asks them to look for differences among the four provinces and differences between them and the pupils' own neighborhood.

Lesson 8 - The teacher begins with a review of the film - asking people to comment on what they saw. He then asks them to make comparisons to draw out any geographical significance from what they saw.

> "Yes, most of the houses in Saint John are made of wood. Why?"
> "What freight might the ferry from North Sydney to Porte Aux Basques be carrying?"
> "Why are trees taller in Fredericton than in St. John's?"
> "Why is there a higher proportion of arable land in Prince Edward Island than in New Brunswick?"

Lesson 9 - The teacher gives a mid-unit test on what the class has done so far. This test will count and the class is given several days warning.

What's right with that?

1. The teacher began by assessing what pupils already knew so as not to bore them and waste their time by teaching it again.
2. The teacher gave clear written objectives. The pupils knew what facts they ought to know, what relationships they had to understand, what things they should try to find out.

3. The film was used to assess what the pupils had understood. and failed to understand.
4. The mid-unit test was used partly, but not entirely, for instructional purposes. The teacher reviewed the work done on the basis of the test results.
5. Instructional assessment took several forms - short answer quizzes; directed oral questions; probing questions; longer written tests.

The reader's reaction at this point may be, "Well, that's all common sense. Any good teacher would do most of those things."

That would be a fair comment. Instructional assessment is common sense and most good teachers have always found out what pupils did not know, filled the gaps, tested, and reviewed. However, the reason I am going into some detail on instructional assessment is that, although it is not controversial, everyone agrees it is important, it is very commonly carried out very badly. The following examples are all too familiar:

1. Teachers sometimes begin teaching without checking to see who already knows what. Time spent teaching what is already known is time wasted.

   Alternative (i) Those who already know the basic facts might be given a different assignment that will help develop excellence.

   Alternative (ii) Those who already know the basic facts could drill the other students with oral quizzes.

2. Teachers sometimes give tests after little and ineffective teaching and count the results even though the tests have been developed for instructional purposes. Such tests often merely reveal that most pupils have, as yet, learnt very little. Instead of providing an incentive for future learning, such failure may only reinforce a sense of despair.

3. Teachers sometimes fail to distinguish between valuable and trivial knowledge, e.g. between knowing where St. John's is on a map and knowing how many kilos of cod were landed in Nova Scotia in 1979. (It might be useful to know how important fish is to the Nova Scotian economy, and how important cod is compared with other products, but certainly it is hardly useful for people to memorize the precise total of cod caught in a single year.)

In determining what is useful and what is trivial, the following are useful questions for the teacher to consider:

(a) in the early grades, "What are the basic skills and knowledge that every functionally literate person must have?"
(b) in the middle grades, "What do we want an educated member of the public to know and understand about, for example, the geography of the Atlantic Provinces?"
(c) and in the senior grades, "What should be expected in advance from a university student beginning a university course in, say, Canadian geography?"

4. Correct answers from a handful of students are frequently interpreted as indicators of general knowledge by most of the students.

5. Failure of many pupils to answer test questions correctly is often interpreted as meaning that the questions are too difficult or that the pupils are lazy. Another possible explanation is that the class has not been directly taught what it is that is being tested.

It would be a mistake to leave instructional assessment without stressing the point that it is not only of interest to the teacher. Learners should be even more concerned about learning than is the teacher. Instructional assessment is important to them because learners should be told exactly what is their problem, and what, if anything, they can do about it.

For example, one sometimes finds a bright child who is reading at or only just above grade level. Teacher tests may show even lower levels of performance, perhaps well below average. A few questions will often show that the child scarcely ever reads outside school. If the school has some kind of free reading program, this pupil is likely to spend that time reading a comic or thumbing through a magazine, neither of which activities is likely to remediate this or any other problem. The pupil should be given a prescriptive list of books that are likely to be of interest and which are a little above the pupil's current reading level.

If learners can be shown a ladder of learning and their own place on it, it is very probable that they will do something to improve themselves. Teachers are sometimes amazed at how even fifteen year olds will scurry around to improve their score in spelling or to improve their time in a fitness test. I believe the reason for all this activity is more than just a natural love of competition--most learners are really satisfied when they can see their own improvement step by step, day by day.

Assessment as an Incentive

Nearly all teachers evaluate for instructional purposes. Almost as many use assessment as an incentive. Nothing concentrates the mind more than the knowledge that one is to be hanged in a fortnight. It is fashionable in some circles to denigrate the use of marks and tests as a means of encouraging pupils to learn. However, one of the reasons most of us work is that we want to eat. Survival is not to be sneezed at. One hears sometimes that extrinsic incentives are bad and that children should come to learn simply for the love of learning. Now it is nice if we like our work and some of us, and some pupils, are fortunate enough to enjoy most of what we do most of the time. Even so, in my experience, even the most committed workers seldom forget to pick up their pay cheques.

In any case, I do not accept that tests are only extrinsic incentives. If the reward for work accomplished consists of money or candy it is probably extrinsic. However, marks in the form of a range of numbers or letters, denote the teacher's understanding of the quality of the submitted work. If the extrinsic reward from work is related to the pleasure of producing something of quality, it can often be legitimated or even induced by the recognition of the established authority. Critics of extrinsic rewards are unlikely to object to a teacher smiling at a pupil in encouragement, or giving the pupil words of praise. It is difficult to see exactly how the giving or withholding of such praise differs qualitatively from the giving or withholding of marks.

Tests and examinations not only provide a general incentive to work. They can also form the basis for some of the most important aspects of schooling. If teachers are interested in the lasting influences they have on learners, rather than on trivial facts learned today and forgotten tomorrow, then the analysis and synthesis of a unit or course are of vital importance. The examination should be the occasion when the student considers the course objectives and puts together the pieces of learning that will satisfy those objectives.

The fact that evaluation, and the resulting marks, form a powerful incentive does not mean that every piece of evaluation should provide such an incentive. For various good reasons, teachers do not ask children to prepare for most norm-referenced test. Such types of assessment have important purposes, but an incentive to learning by students is often not one of them. Such tests may still form important incentives to teachers, who may receive some of their intrinsic rewards from the academic success of their pupils, but that is not the topic of this book.

If evaluation is to serve usefully as an incentive, the testing must be closely related to legitimate objectives, and that connection must be appreciated by the learners. English teachers sometimes give occasional spot tests without warning. Their purpose is to ensure that students read the assigned chapters or passages regularly. The spot test is fairly effective for that purpose but it is frequently counter-productive for other purposes. The easiest way to find out whether pupils have read a chapter is to ask a few true/false or open-ended information questions about salient facts. To make sure the students do not get summaries from friends just before the class begins, teachers will often put in a few obscure questions such as, "What color was the hat Miss Grandy was wearing when she went to church?" That kind of question, if we assume the color of the hat is insignificant, can inhibit and discourage intelligent reading and encourage the idea that reading has no other purpose than the compilation of trivia. The same principle applies to a spot test in a chapter in geography or history.

One way around this is to confine spot tests to questions of major purpose and significant detail. Another is to alternate factual, spot tests with spot tests of a more searching nature -- appraisal of the chapter; analysis of character development in the chapter; the linkage of the chapter with earlier sections. It is true that such searching tests mean much more work for the teacher. However, they also serve broader purposes and yet still serve as incentives for students to read assignments. If students do not know which chapter is to be subject to this treatment, they may be encouraged to read every chapter systematically in terms of the teacher's objectives for the work in progress. On the other hand, if the chances of being tested become rather remote, the students, particularly sophisticated secondary students, may decide to take the gamble. The result could be a set of invalid results, with a random sample of students being totally unprepared and therefore scoring zero. If this is a problem, there is really no alternative to testing every section.

So far, the discussion has been principally in terms of positive incentives--incentives to analyze, synthesize, appreciate. However, the reader may well be thinking, "Well, that all sounds very nice. But the truth is that the real incentive of school testing is a negative one--'If you flunk this test, kid, watch out for trouble.' And surely negative incentives are bad things."

The response to that point is that teachers use both positive and negative incentives, and the two are frequently inextricable and complementary. Desire for excellence in one learner may be coupled with fear of not achieving it. Another learner facing the same test may simple be influenced by fear of failure. There is nothing wrong with negative incentives, within reason. We are all influenced daily by negative as well as positive incentives. Certainly, if the only incentives we experience are negative the outcomes are likely to be unfavorable. Certainly, positive incentives are used far too infrequently by many teachers, but let us not pretend that fear of failure can or should be excluded from the school. After all, as the next two purposes of evaluation clearly demonstrate, schooling does matter. Any pretence by teachers that progress in school does not really matter is either harmfully ignorant or fraudulent.

Assessment as a Means of Classifying Learners

Probably, the single most important function of evaluation in the school is the classification of pupils. They are typically graded on letter or percentage scales. Some critics argue that schools should get out of the classification system altogether. I shall briefly outline the reasons why classification is both desirable and necessary, recognizing that these reasons will not convince the determined critic.

1. Children want to know where they are, as do their parents; progress in school has a major influence on future job, income and social status; it is

unfair to hold back such vitally important information on academic progress until it is too late to do very much about it;

2. Children enjoy competing well with one another to achieve higher standards--this is a corollary of the incentive argument;
3. Competition is a natural part of our society and only a political decision to move to a much more egalitarian, communist society could lead to the abolition of competition in schools;
4. One of the purposes of education is the encouragement of quality and excellence; if they are never recognized, it is difficult or impossible to encourage them; how can one encourage what one does not identify and reward?

The reader may or may not be convinced by those arguments. The fact is that nearly all schools use some kind of classification system. The most important quality of assessment intended for classification purposes is justice -- the test must be fair. One of the most important criteria by which students judge teachers is fairness and that fact in itself illustrates the importance of classification in our school system today.

Teachers are sometimes quite surprised and upset because pupils feel they are being treated unfairly. Teachers frequently interpret such complaints as meaning that pupils are only interested in marks. They may be at least partly right--bear in mind that marks are critically important in the life of the school. Most of the traditional control mechanisms have, rightly or wrongly, been removed from the school. Corporal punishment is rare. Suspension and expulsion are difficult to accomplish. Teachers rarely receive the automatic, traditional, unquestioning respect they once did. Many large urban schools have little sense of community--the pupils have little guilt about "acting out" or being destructive. Such behavior will not be the subject of gossip and criticism as it is likely to be in a more closely knit rural community. Thus marks, which are critically important in their own right in terms of the pupil's future, become one of the school's very few remaining mechanisms of control. While I am not arguing that young people should only be interested in marks, and should not be at all interested in learning, I do think that teachers are being disingenuous when they affect an attitude of disdain towards evaluation. The student who disarms the authority of the school by declaring a lack of interest in marks is likely to be a danger to the organization.

If evaluation is to be seen as being fair, students must understand the classification system. One of the objections to classification is that it is demeaning to those who do badly. The problem then is to develop a plan which will classify children accurately, so they and their parents know where they are in their learning, and which also gives everyone a chance to make reasonable progress and avoid being stigmatized by failure. The incidence of failure can be reduced by good instruction and intelligent placement. The consequences of failure can be mitigated, but not eliminated, by carefully developed grading schemes. That said, failure remains an integral part of any classification scheme for grading pupils. There is no success without failure, the word itself losing sense outside that context. This remains true whatever labels are attached to the system of classification. In some schools numbers are used without cutting points. There is no stated level of failure. Sometimes, letter-grades 'A', 'B', 'C', 'D', & 'E' are used but no 'F'. Others will simply use categories of satisfactory ('S') and unsatisfactory ('U'). Whatever the euphemism, failure remains a concept well understood by pupils, even in the early grades. Reducing the number of classification points to two only serves to emphasize the disgrace of being "Unsatisfactory" and to remove any indication of and incentive for excellence. Individual schools will naturally vary the emphases they put on comparative success. No contemporary school can eliminate completely implicit or explicit comparison. I am not arguing here for an extreme position. I am not arguing for public humiliation and the castigation of failure. I am saying that relative degrees of progress should be clearly and calmly made available in an open manner. I am also recognizing that failure cannot, and should not, be totally eliminated from a classification scheme.

Assessment for Certification

Pupils are classified in school almost daily, in reading, chemistry, and even behavior and attitudes. Certification is the logical outcome of the classification procedures. Children are certified as having completed Grade 1 or senior public school, and high school. The function of certification is a little different conceptually from that of classification. Classification is essentially comparative. Students are given relative grades which lose their value if everybody receives the same results. Certification has a somewhat different emphasis. It is possible, on occasion, for everyone to graduate. Essentially, certification is supposed to be a stamp of competence. If everyone reaches the minimal level of competence, then everyone should receive the just reward. Unfortunately, to an increasing extent, certificates of competence (graduation certificates, degrees, etc.) bear little relationship to specific standards. Indeed there is even pressure to distribute them evenly among different groups in the population! This trend is unfortunate. The consequences are declining faith in the credential provided by educational institutions on the part of the employers and the public. The loss of faith could lead to the development of tests by employers and student disaffection with schools and universities. Make no mistake about it, most young people want employment opportunities as well as a broad education from their schooling, and certification is vital for this purpose.

Certification is an important function of the school precisely because it provides a basis for the allocation of future training, income, employment and social status, and these are important matters in a free, democratic, individualistic, and materialistic society. It is sometimes argued that it would be better if schools were not party to the process. But if certification does not determine employment and social status, what will? Those who object must provide an alternative.

If the job has to be done, it should be done well, and well means fairly. That does not imply that the school should emphasize the meritocratic, certifying aspect of schooling and ignore its broader philosophical framework. Young people should learn to enjoy aesthetic experiences, to be physically fit and to act upon a coherent set of acceptable moral values as well as get themselves certified. The argument here is only that the certification process, which the schools carry out, should be accomplished openly, fairly, and efficiently and not that it should override all other objectives of education. Some will say that schools should get out of the certification business and leave the problem of selection to employers and universities. That is certainly an alternative model and one that should be considered within the normal democratic and political process. I shall not here discuss the implications that such a choice would have for the concept of compulsory, public education. Suffice it to say that the certification function is currently the job of the school and that there are strong reasons why it should remain so.

Classification and certification are closely related but each emphasizes a different aspect of assessment. Classification implies comparing students with one another or with some explicit or implicit norm. Certification implies assessing students in terms of a promulgated standard of competence. The distinction is not a rigid one. Classification may often assume an overall standard of quality. Certification is dependent on the development of realistic, one might almost say, normative levels of competence. It would make as little sense to classify a student as being 'A' in mathematics just because he is the best of a lazy, incompetent group as it would be to fail to certify as doctors all medical students in a large class on the grounds that none meets our new standards of brilliance, excellence, and moral probity. Nevertheless, students are generally classified in terms of one another, and certified according to some standard.

Summary

This chapter has been concerned with the basic purposes and functions of the evaluation of students in school. Instructional assessment is an essential part of the teaching operation. Increasingly it is recognized that effective teaching is dependent on accurate assessment. Such assessment can be carried out formally and informally. The teacher should be careful not to count results of tests administered for instructional purposes as part of pupils' record of achievement.

Evaluation is often effectively used as an *incentive* to work productively. Sometimes the incentive may be counter-productive, if, for example, pupils learn meaningless mumbo-jumbo because it will be "on the test." A good test will assess good objectives sensibly, and preparation for that test should be useful. A good test will provide an incentive to pupils to learn effectively and, possibly, synthesize what they have covered in the unit, term or year. Evaluation also serves as a powerful negative incentive in the form of fear of failure.

Evaluation is used to *classify* learners both in terms of the level at which they are working and in terms of the relative progress they are making compared with others working towards the same objectives. If classification is not to countermand the incentive evaluation provides, it must be used to reinforce excellence without providing continuous negative reinforcement to the less successful. This can best be achieved through a dual system of classification. The first part classifies the learner in terms of the appropriate level at which he or she is working. The second part classifies the learner in terms of the progress the individual is making at that particular level. In this way it is possible to meet the instructional criteria of giving every learner situations where success can be achieved without sacrificing the necessities for classification that demand a universal system based on clear, accurate, public information.

Classification leads to *certification*. One of the functions of schools in a meritocratic society (and all developed countries, communist and non-communist, are, to a greater or lesser extent, meritocratic) is the allocation of opportunities for future social advancement. It is essential that evaluation of learners be based on valid and reliable measures so that justice can be done and can be seen to be done. This requires a great deal of openness in terms of objectives, testing, marking, weighting, and reporting.

2 TESTS

Their Characteristics, Use, and Abuse

This chapter is intended to give an overview of tests and their characteristics. Emphasis is on tests that are likely to be of direct interest and use to the classroom teacher. No attempt is made to cover the huge variety of tests that is available--particularly tests used for the measurement of individuals and their special characteristics. Instead, the emphasis is on the problems and techniques of greatest interest and value to the typical classroom teacher.

Validity and Reliability

In general, tests should be both valid and reliable. A *valid* test is one that measures what it is meant to measure. For example, suppose a teacher gives a list of words to a class and asks that they be memorized. The following day Maria gets her list entirely right whereas Brian gets several words wrong. Let us suppose the pattern repeats itself over several months. The teacher may conclude that Maria is a better speller than Brian. However, tests of lists of memorized words are not necessarily totally valid measures of general ability in spelling. It may well be that Maria is extremely conscientious but is nevertheless a bad speller in usual practice. Perhaps she has a small vocabulary and may make frequent errors even within that small vocabulary when she actually writes. Brian may make as many errors with a small standard vocabulary but he may be more proficient when he exercises his larger vocabulary.

I am not saying that memorization of spelling lists is necessarily an inappropriate teaching technique, though I would question a very heavy reliance on it precisely for the reason indicated above. The weekly spelling test may well provide a good incentive to learn, and may provide the teacher with a good means of discovering problems. However, by itself, it is unlikely to be a valid means of classifying a person as a good speller.

It is often difficult to determine how valid one's measures of performance are. For example, the English teacher is understandably apt to rely on written composition and answers to comprehension questions as a means of classifying young learners' competence in written English and reading comprehension. If there is much reliance on homework however, real problems of validity arise. Was the work done alone or did parents, siblings, and friends help? If the work is done inadequately (or not at all), is that a measure of incompetence or merely a measurement of unwillingness? Once again, I am not suggesting that the homework assignment is an inappropriate means of evaluation and instruction. I am suggesting that heavy reliance on it for purposes of classification or certification is unlikely to be valid, if the claim is being made that the marks reflect competence in the cognitive skills in question. It may be argued that industriousness, obedience, punctuality, and independence are important traits which the school should be teaching. Marking English based on homework assignments may indeed be a reasonably valid measure of both a combination of those traits and competence in written English and reading comprehension, if the assignments are completed independently. However, it may be better to separate the evaluation of such rather different characteristics.

The kind of validity discussed so far is often known as face validity. Thus a test has face validity if, upon examination, it appears to experts to be measuring the objectives one wants to measure. One can measure *concurrent validity* by checking to see if other measures of the same concept arrive at similar results. One can measure *predictive validity* by finding out whether results on tests indicate behavior in the future. Thus a measure of the predictive validity of a university entrance examination might consist of a study to see if students who did better on the examination perform better in university.

A *reliable* test is one that will give the same result in different circumstances. For example, if a set of tests can be marked independently by various competent teachers to produce the same results, it is reliable. Thus good objective tests tend to be more reliable than are essay or paragraph assignments. If a class is given two comparable tests under the same conditions at about the same time, the results should be identical if the tests are reliable. Essentially then, a reliable measure is an accurate one. Unfortunately, it is not always easy to judge just how accurate a particular test will be in a particular set of circumstances. A reliability measure of 0.85 sounds quite good but the statistic may have been obtained in conditions very different from those in which a teacher plans to use the particular test.

There are several ways in which the reliability of an apparently reliable test can be affected. Firstly, the testing conditions may be inappropriate. They may be inappropriate because the pupils have been told that the test results do not matter. If the results do not appear to matter, it is not surprising that some pupils make little effort. One can go to an extreme on test conditions--and I have heard some teachers claim that you should never test on Monday, Friday, early morning, or later afternoon! Yet those times are used for normal instructional activities. Secondly, tests are unlikely to be reliable if they are inappropriate for the pupils taking them--if they are much too difficult or much too easy. Similarly, as an example that follows shows, if a teacher uses tests on different populations from those on which the reliability coefficient was measured, the reliability experienced by the teacher may be very different. That problem can be approached by looking at the way reliability is assessed.

The reliability of published tests is assessed in a number of different ways. One way is to divide the test into two equivalent halves and rank the students taking the test separately on those two parts. If the rankings are identical, the test is perfectly reliable for that group of students in those conditions. That example raises a problem. If the students are almost identical in terms of achievement, if, for example, all the students have the same ability and have had identical instruction, the chances of the two halves giving identical rankings are nil. The differences between the students are just too fine to be reliably measured. If, on the other hand, the age and competence of the students are extremely varied, the rankings on the two halves are more likely to be similar. For example, if verbal I.Q. scores are taken for a group of children of similar age at the beginning of second grade, they will probably predict the rank order of those children in reading achievement fairly accurately. The children will represent a considerable range of measured ability. However, if the verbal I.Q. scores of university applicants are used to predict performance in first year university English, the prediction is likely to be highly inaccurate.

The range of learning ability is greatly reduced by the time students apply for university entrance. In general, the less able students have dropped out or made other choices, and the few remaining less able students are likely to be sufficiently highly motivated and organized to compensate for lacks in intelligence. Yet both tests might have been given the same reliability coefficients by their makers, based on the samples they used.

There is another aspect of reliability beyond the variation in those being tested. It is usually easier to test a single skill reliably than multiple skills. Suppose we were to test the mathematical proficiency of five individuals--one a university professor of mathematics, one a university student of mathematics, one an eleventh grade student taking mathematics, one ninth grade student and one student who has just failed sixth grade mathematics. If we devise a broad test of mathematical skills, we should be able to predict the results with exact reliability. Suppose, however, we also want to include in our test measures of swimming, running, reading, and moral reasoning. We want to find who is, on average, the best. The results will not only be different, they will also be less reliable. There will be some day-to-day variation in performance in mathematics--and even the mathematics professor may not be very good at rapid computation when she is upset. If we add all the variations in the other areas, the average rank order of our five contestants is likely to vary considerably according to test conditions and other factors. In other words, the greater the variation in the objectives being tested the less reliable the results are likely to be.

Unfortunately, important educational decisions are likely to be made in circumstances that are not conducive to reliability. Suppose a university wants to pick a group of entrance scholars. It is likely to be interested in a wide range of criteria (English, mathematics, science, social science, general intelligence, range of interests), and to be faced by a narrow range of applicants, most of whom will rank high compared with the rest of the population in terms of those qualities. It would be much easier for the university to be reliably fair, if it were interested in only one criterion and were faced with a wide variety of applicants. That, of course, is not the case.

This problem is important to the classroom teacher. The classroom teacher has to evaluate a wide range of skills and yet is faced with a relatively homogeneous group of students. In those circumstances, the reliability of the measurement, whether it is made by a teacher-made test, a standardized test, or any other measure, is unlikely to be very high. Reliability is usually expressed on a scale from 0.0 to 1.0 where 1.0 means perfect reliability--one test predicts perfectly the rank order of the people being tested on another test. A commercial test may claim a reliability of 0.85 or 0.9. Its reliability may not be so high in the circumstances in which the teacher wishes to use it.

Validity, Reliability--Can One Have Both?

It is desirable for tests to be as valid and as reliable as possible. If one has a very straightforward objective (e.g. the testing of addition) and a wide range of learners, one can easily devise an extremely valid and reliable test. The test is not going to be valid if it includes questions on all kinds of other things that have not been taught. It is not going to be reliable if it includes many ridiculously extreme questions, e.g. the addition of twenty 30-digit numbers, and not valid if it includes only questions involving single digit numbers.

Sometimes, however, one has to make a choice between reliability and validity. For example, if one wants to classify students in tenth grade Canadian geography, one could make the test more reliable by confining oneself to a few fairly straightforward objectives. One might restrict oneself entirely to actual, true/false questions. In the face of a broad set of objectives, such a test is not going to be valid. Yet the more one extends one's questions to more objectives, and more complex objectives, the less reliable the results are going to be. It is at this stage that educated judgement becomes very important. Fairness demands some degree of reliability. Pedagogy demands some degree of validity.

A test is likely to be reliable in the following circumstances:

1. The objectives are clear and understood by the learners;
2. The questions are clear and unambiguous;
3. There is no desire to make very sharp differentiations among similar students;
4. The range of objectives being tested is not extremely large and varied.

A test is likely to be valid in the following circumstances:

1. The objectives are clear and well known in advance;
2. The test items are clearly identified with and representative of the objectives.

This discussion should not be taken to suggest that only simple objectives can be reliably and validly measured. Critics can often reach high levels of agreement about the relative competence of concert pianists, even though the criteria are many and complex. Tests of learning ability (I.Q. tests) are often highly reliable, although some critics doubt their validity. Before examining that sensitive topic, it is worth taking a brief look at three other terms used frequently in the field of evaluation.

Diagnostic, Formative, and Summative Evaluation

Diagnostic evaluation is intended to find out exactly what learners know and understand. They are sometimes used by teachers before instruction begins. They are also used during instruction as part of a scheme of *formative* evaluation. Formative evaluation consists of the testing and assessment of learners carried out during the instructional process. In contrast, *summative* evaluation is testing and assessment carried out at the end of a unit, term or year. Formative evaluation is usually intended to diagnose problems and to provide an incentive for continuous learning. Summative evaluation is usually intended to provide an incentive for learners to synthesize what they have learned and to classify students relative to one another. Sometimes it also has a certifying function, e.g. to denote high school graduation. Summative evaluation can also be used for diagnostic purposes, although this possibility is frequently overlooked. Even at the end of a unit, it is useful to find out the degree to which different objectives have been achieved. Summative evaluation provides an excellent basis for program evaluation.

What Exactly Is Meant by Diagnosis?

I do not wish to weary the reader with a lengthy semantic discussion about diagnosis. We educators use the word in a number of different ways and thus confuse one another as much as we do parents. Sometimes we use it to describe assessment for instructional purposes simply to indicate testing that finds out what children do or do not know or understand. However, it is often used at a deeper level, to refer to the discovery of deep-seated problems, and perhaps even to include possible remedies.

Before turning to the I.Q. test controversy, it is useful to distinguish between achievement tests, which test students' performance on task, and tests such as I.Q. tests which purport to test students' potential abilities in one way or another. If we apply the second meaning of the word diagnosis, the I.Q. test becomes a diagnostic test. Generally speaking, teachers rightly put more reliance on measures of actual performance, which they can seek to improve in some direct way, than on the more remote measures of ability, readiness, or aptitude, which are at least one stage removed from performance. Certainly there are dangers in jumping to educational conclusions from diagnoses of varying reliability--of dyslexia, personality disorder, or "lack of readiness." Learning ability, however, is something of a special case as I.Q. scores are not only highly reliable, but are both predictors of future performance and, according to many, but certainly not all, researchers, indicators of genuine intellectual potential.

Learning Ability Tests (I.Q. Tests)

I.Q. tests are probably the most controversial of all tests used in schools. Indeed, some Canadian school systems and many American jurisdictions have banned their use altogether. In general, they are less fashionable in Canadian schools than they once were.

The essential principle of the I.Q. test, at least as administered to children up to the age of 14 or 15, is that a comparison is made between the so-called mental age and the chronological age. I.Q. tests are scored on a normal curve with a mean of 100 and standard deviation (by which I mean, approximately, the average variation from the average) of between 12 and 15. One assumption is that there is a normal rate of maturation and that there is some stability in the degree of rate of growth within an individual. In other words, children with I.Q.s greater than 110 at age 5 are assumed to have I.Q.s above average at ages 10 and 15.

People with I.Q.s below 70-75 are usually considered mentally retarded and in need of special, usually segregated education. Those in the 70-90 range are considered slow learners. The normal range is between 85-90 and 110-115. An I.Q. over 120 is usually considered an indication of superior intelligence and lower limits of 135 to 150 are used to identify the intellectually gifted.

Most university students in Canada have I.Q.s over 100. An I.Q. below 90 is a predictor of severe difficulty in an academic high school program. Young people of superior intelligence are likely to be able to cope intellectually with almost any university program.

The Case against the Use of I.Q. Tests

1. They are unreliable. Individuals' scores do vary from time to time, and certainly from year to year. Furthermore, a child who is nervous or upset may do particularly badly.
2. They are invalid. It is argued that there is no such thing as general overall intelligence. Intelligence is principally acquired through environmental experiences. Therefore, an I.Q. test is at best an inaccurate measure of some of the experiences an individual has already undergone. Educators should be concerned with future experiences. Therefore, their only need is to know the degree of understanding of necessary prerequisites for the achievement of subsequent objectives. Intelligence, whatever it is, is irrelevant. To the extent that there is inherited intelligence, it consists of a set of discrete abilities that cannot be captured in a single score.
3. They are discriminatory. Children from low socio-economic backgrounds score lower on I.Q. tests than do children from more affluent homes. Children whose first language is other than that in which the test is administered are likely to do poorly. I.Q. tests are culture bound in that they are based on samples from "normal" homes exhibiting "normal" upbringing. This is disadvantageous to children who do not reflect such norms. I.Q. scores developed in such questionable ways can be used as a basis for discriminatory treatment and segregation.
4. They are open to abuse. Once an I.Q. score is taken and recorded, it can be used as an explanation of why a child does not learn. It can thus become a self-fulfilling prophecy. Children are classified as being slow, treated as being slow, and then become slow.

The Case for the Use of I.Q. Tests

1. I.Q. tests, even group tests, are fairly reliable measures of general intellectual functioning of most young people in normal circumstances. They are also valid, providing the people being tested come from the cultural and linguistic backgrounds for whom the tests were developed. Obviously, there is no more point in giving an I.Q. test in English to a unilingual Italian than there is giving a teacher's test in Latin to a student who has never taken that particular subject.
2. There are, in all probability, a number of aspects of intelligence. We do not know very exactly the extent to which such qualities as numerical and analytical ability, verbal reasoning, spatial ability, and imagination and creativity are discrete or associated properties. We do not know the exact degree to which those different aspects of intelligence stem from genetic inheritance as compared with the cultural environment. We do know that I.Q. tests measure numerical, analytical, and verbal abilities most, certainly more than they do imagination and creativity. Nevertheless, there does seem to be a generalized intelligence factor that differentiates people in terms of their ability to carry out a wide range of intellectual tasks. In study after study, I.Q. proves to be the single, best predictor of performance on virtually any academic task. Some psychologists do claim that "there is no such thing as intelligence." By careful semantic manipulation one could maintain equally well that there is no such thing as peace or happiness. Nevertheless, intelligence is a concept most teachers have little difficulty understanding. Quite young children differentiate between the "smart" kid and the kid who does well because he is the teacher's pet or because he works hard. A thirty-minute physical examination pronounces one healthy. A thirty-minute test pronounces one intelligent. Neither evaluation is infallible but both are valid within their limitations. It is more likely that the man pronounced healthy will drop dead of a heart attack within a few weeks than a child pronounced a genius will turn out to be mentally retarded.
3. I.Q. tests, like all measures, discriminate. Discrimination is the basis and purpose of evaluation. Children from low socio-economic backgrounds do score lower on the average on I.Q. tests than do children

from high socio-economic backgrounds. Such children also score lower on reading, mathematics, and all other tests of achievement. When I.Q. tests were discarded as a basis for selection in one English school system in favor of teachers' recommendations, the system found that teachers identified fewer able children from poor homes than had the tests. In other words, the teachers were, apparently, more unfairly discriminatory than were the tests. There remains the question: does the I.Q. test discriminate *unfairly?* That brings us to the lengthy debate about nurture and nature. Does I.Q. come from genetic inheritance or from cultural upbringing? The answer is both. If I.Q. scores stem principally from culture, it can be argued that their use predisposes educators to have unfairly low expectations of children from poor homes. However, the best evidence is that I.Q. comes neither less than 50% nor more than 80% from inheritance and probably not more than 20 to 30% from cultural upbringing. Nutritional patterns before and immediately after birth, together with other comparable experiences, probably account for additional, minor components of I.Q. variance. (It is important to note I am talking about averages here; a woman who drinks and smokes heavily during pregnancy may well be doing catastrophic damage to the I.Q. of her unborn child.) One remarkable thing about I.Q. scores, to many educators who have been working with them for a number of years, is the frequency with which children from the poorest homes score remarkably well on I.Q. tests. This constitutes one of the strongest reasons for continuing their use. Far from condemning poor children to poor education, I.Q. tests help identify bright children who might otherwise be considered slow.

4. I.Q. test scores can be put to good pedagogical use. The most obvious use is in the case of the talented individual from the poor home. Unless that talent is noticed, nurtured, and encouraged, it is unlikely to be productive. Just as teachers sometimes fail to recognize bright children, so can parents, particularly if they themselves are uneducated. Associated with this use is the identification of children achieving far below the level of which they are capable. Quite bright children sometimes are left to plod along slowly on their own, particularly if they

are not trouble-makers. They may be categorized as being slow, and slow they are if we are referring to their rate of achievement. But slow is often interpreted as being unintelligent by teachers, parents, and the child himself.

I.Q. scores can also be useful, on occasion, in the case of a child that works too hard. Parental pressures, a determination to succeed, and the success of siblings may combine to make school a hell for such children. Sensitively handled, a low or moderate I.Q. score may help persuade parents to lift undue pressure and find other outlets for the child's energy and persistence. One obviously has to be much more careful about lowering expectations than about raising them but there are occasions when parents hold totally unrealistic hopes for their child.

I.Q. scores are also useful in determining whether or not one should accelerate young children. Acceptance of acceleration does not imply a desire to push children through school as fast as possible. It does raise the possibility that the range of ability and competence in the elementary school is so great that bright children may be held back and bored. This problem is explored more fully in Chapter IV.

In sum, I.Q. scores, handled sensitively and intelligently, can be very useful. They should never be used independently to make major educational decisions but they should be used in conjunction with other data--achievement test results, teachers' marks, teachers' assessments, and observations. Two or three similar scores, e.g. 107, 114, 109, probably give a fairly reliable and valid picture of what the average teacher means by the term intelligence. Given that cultural background certainly influences intelligence to some degree, I.Q. scores tend to understate somewhat the "true" or genetic intelligence of children from poor homes. Perhaps five to 10 I.Q. points may be explicable in this way. I.Q. scores are particularly useful--in determining the kind of help most suited to those with learning problems, e.g. placement in a class for the mentally retarded as compared with withdrawal for special help in reading and mathematics in a remedial class for children of normal and above learning ability; in determining whether a high achiever should be accelerated; in selecting the gifted; in determining whether to challenge and push an apparently bright child by placing her in an advanced reading group--even though her achievement is not much above average; and in counselling parents of an overly pressured and abnormally conscientious child who is showing signs of emotional problems apparently as a result of the family background.

Standardized Tests of Achievement

The most commonly used standardized tests are norm-referenced tests of academic achievement. Such tests are most often used to assess progress in the basic skills, particularly mathematics, reading and language.

Norm-referenced tests provide scores around a norm. Thus if a random sample of fifth grade children across Canada were to take a reading test in December of fifth grade the average score would be set at 5.3 grade levels. The variation in reading skills tends to be considerably greater than is the case with mathematics. Thus, fifth grade students quite frequently score above the ninth grade level in reading but it would be rare for such students to score above the seventh grade level in mathematics. One must be cautious therefore in saying that a student is "better" in reading than in math "because the grade level is higher." It is very difficult to make comparisons about relative performance on two entirely different tests. Is the achievement spread really greater in reading than in mathematics, or is it just that improvement in mathematics is more dependent on specific instruction? Ironically, teachers tend to mark in the reverse order. Math teachers tend to use a wide spread of marks, English teachers a narrower spread. I shall return to that problem in Chapter III. At this point, I merely want to emphasize that tests are man-made and can be constructed with many very different characteristics.

Many people prefer to report norm-reference test results in percentiles, deciles, or quintiles, partly because there is uncertainty about exactly what "reading at a ninth grade level" means. If percentiles are used, the mean becomes the 50th percentile and a student in the 90th percentile has performed better than 90 per cent of the students on whom the test was normed. A student in the 14th percentile has performed better than 14 per cent of the students in that same group. Deciles are percentiles arranged in a 10-point spread--the first decile being from the 90th to the 99th percentile. Quintiles are percentiles arranged in groups of 20, so the first quintile consists of those students above the 80th percentile and the fifth quintile consists of those below the 20th percentile.

If a choice is possible, I would recommend the use of quintiles for the presentation and recording of standardized test data. Percentiles are an obvious alternative, but they can be confused with percentages. Grade levels are quite useful in the elementary grades but there is always the danger that a parent will believe that the child of ten with a score of Grade 6.8 should be placed automatically in sixth grade. Parents may also be upset if the child who scores 9.4 at the beginning of grade 5 only scores 9.3 at the end. In fact, it is most unlikely that a standardized test will be very reliable at the extremes. Unfortunately, one must also admit that it is almost as unlikely that a high achieving fifth grader will have benefited measurably from typical fifth grade reading instruction.

The Case against Norm-Referenced Tests

1. Norm-referenced tests tend to evaluate a rather narrow range of objectives in order to maintain high reliability. Thus one should be cautious about using the results to make overall comments on a student's academic progress. The objectives tend to be the most straightforward and most easily quantified, and, some would say, the most trivial.
2. As results from the norm-referenced test have an aura of scientific truth, they are likely to be given a high level of credence by the public. There is danger that teachers will teach to the test, ignoring other more important, untested objectives. For example, written English may be neglected in favor of grammar or language usage.
3. Although norm-referenced tests are highly reliable, they are not the best predictors of future pupil performance. The best predictor of performance in grade eight English is performance in grade seven English. This fact makes some people doubt the validity of such tests.

4. Norm-referenced tests, by definition, assume a national norm by which all are judged. Some educators believe it is harmful to judge all children by the same criteria. It builds up an expectation, they believe, that all children should learn the same amount at the same rate.

The Case for Norm-Referenced Tests

1. Norm-referenced tests can usefully be used to measure the achievement of those objectives most educators hold in common. Those objectives, including reading comprehension and mathematical skills, are rightly seen as basic skills of central importance.
2. It is sensible for teachers to teach to the test in the sense that good tests should measure good objectives. The achievement of all objectives should be evaluated regularly. Reading and mathematics are objectives of central importance and they should certainly form part of the evaluation scheme.
3. Norm-referenced tests predict test results and teachers' marks predict teachers' marks. Teachers look for different things from tests (obedience, conformity, neatness, dependability), so it is not surprising that teachers' marks correlate with other teachers' marks. Norm-referenced results are important because they provide the public with a score related to the nation, whereas teachers' marks can only be interpreted in terms of the rest of the class, or at best, the teachers' prior experience. A teacher can teach very badly and still give high marks. Standardized tests provide a check. The situation is comparable to that of the doctor. Doctors usually make good diagnoses but it is frequently desirable to have additional laboratory tests as a check. Professional doctors make good use of tests to complement their own skills. Professional teachers do likewise.
4. Norm-referenced tests should be used as part of an evaluation scheme, not as the only measure. It is perfectly possible, and probably desirable, to report pupil progress in terms of the group in which the individual is working. Nevertheless, it would be unprofessional not

to inform parents, and children, of the level at which the child is working. This information is most accurately obtained from norm-referenced results. One isolated teacher cannot possibly make universal assessments without resort to data from outside the school.

In sum, norm-referenced tests are most useful in tracing the progress of individuals and of groups of students through the school. Test results, modified by experience and other information, provide a good basis for the grouping of pupils for instruction. Different teachers, of course, have different standards and that makes their marks difficult to use for grouping purposes.

Neither norm-referenced tests nor teacher tests measure all the objectives of a school program. Norm-referenced tests should be used within an overall framework of evaluation. It is unlikely that pupils will give them undue attention, unless encouraged to do so by adults. After all, the norm-referenced test does not usually "count" on the home report.

Criterion-Referenced Tests

Whereas norm-referenced tests are designed to compare the performance of individuals or groups with a national norm, criterion-referenced tests are designed to assess pupils' achievement of specific tasks. The differences between the two principles are often exaggerated. A norm-referenced test necessarily includes specific tasks on which individuals are judged. Similarly, some normative judgement has to be made in selecting specific criteria for use in conjunction with criterion-referenced tests.

In practice, published criterion-referenced tests tend to emphasize rather narrow specific objectives. Pupils are assessed, for example, in terms of mastery or non-mastery of numerical computation. Criterion-referenced tests are often very reliable, but it would be rare indeed for a battery of criterion-referenced tests to cover a significant proportion of the objectives of a school program. Some objectives simply do not lend themselves to assessment in terms of mastery. Some are difficult to define in very clear behavioral terms. I refer here not only to objectives in the areas of values and attitudes but also to the more complex intellectual objectives. For example, criterion-referenced tests, as currently developed, are unable to deal acceptably with an objective such as, "the student will be able to synthesize important ideas from a reading list of significant plays and novels."

Teachers' own tests are frequently a modified kind of criterion-referenced test. Published tests obviously cannot reflect the various different ideas and materials used by teachers. Teachers' tests are often both criterion-referenced (in a very general sense) and diagnostic in nature. Other teacher made tests, particularly end of term examinations, are designed more to rank learners on a scale, perhaps from A to F or from 100 to 0. Such tests are more comparable with norm-referenced tests in their purpose.

Table 1

Some Test Types and School Uses

	Formative	Summative
Norm-Referenced	Standardized tests for group placement	End of year or term examinations; standardized norm-referenced tests for records and reports to parents.
Criterion-Referenced	Teachers' diagnostic tests	Teachers' tests to report overall weaknesses and strengths

Standardized Individual Tests

The previous discussion has concerned group tests, i.e. tests designed for use with a group (not, as is sometimes mistakenly believed, designed to give scores about groups, although they do also provide that service). Teachers are rarely concerned directly with individual tests, which are usually administered by registered psychologists.

Teachers frequently treat group test results with a healthy scepticism. Perhaps they see them as a threatening alternative authority. In contrast, individual test results come with all the mystique of the psychologist and can easily be given too much credence.

This book is not the place, and I am not the writer, to discuss the values of the various kinds of individual tests. The most commonly used type is the individual I.Q. test--a somewhat more reliable version of the group I.Q. test. The comments made about the group versions are generally applicable here too. Tests designed to measure various learning disabilities such as poor sight, hearing and coordination are important sources of information for the teacher. However, the single most important piece of information, sometimes provided by the educational psychologist, is an exact description of the learning competencies and inadequacies of the individual pupil. There are usually disagreements among the experts about which exactly is the best method to apply to a particular disability. Yet virtually every expert will agree that one has to start from the actual learning base of the individual child.

Item Banks

An item bank, in its simplest form, is nothing more than a collection of test items from which educators can draw. If a province were to pursue the idea of the item bank to its logical conclusion, it would develop a set of items to match every single educational objective from kindergarten to Grade 12. It might categorize items on a scale of levels of learning (perhaps from primary to university entrance) and also by level of difficulty within the learning levels. There might be a key word index for a teacher to check and draw items according to a variety of descriptive categories. In its most sophisticated form, which does not yet exist, a teacher in Grade 3 would be able to draw a set of items for his class, in, say, reading, made up according to his own choice of objectives. The students' answer sheets would be fed into a computer that would immediately provide every student's personal score together with an indication of the student's score in terms of her peers in the province and in terms of her peers of the same learning level.

Relatively sophisticated item banks have been developed in several countries as a result of real and alleged inadequacies of teacher-made and commercial tests. Criticisms of teachers' tests are that they are frequently unreliable. They deal mainly with low level knowledge objectives, and their standards vary considerably. Criticisms of published tests are that they have a restrictive effect on the curriculum.

In Canada, Ontario has moved into item banking on a large scale with its Ontario Assessment Instrument Pool. Items in Ontario's banks are accompanied by a p value, or some other indicator of level of difficulty. The p value is the percentage of learners in a particular category who answer the item correctly. Those learners are usually a random sample from a specific population--such as 8th grade Ontario students who have or have not been specifically instructed in the topic being evaluated; or 10th grade students taking advanced English.

The most basic and important use of the item bank is the help it provides the classroom teacher. The difficulty of developing good tests is well known. Therefore, a bank of items covering all objectives in a course of study is most useful to the teacher. An example of a way in which a classroom teacher may profitably use item banks is in the construction of two similar tests for application before and after instruction. The teacher might collect two sets of twenty or so items testing the same objectives and then arrange the two sets fairly--so that the mean "p" value (or difficulty level) of the two sets is equivalent (as in Table II). One hopes students perform better after instruction than before.

Table II

Constructing Comparable Tests from an Item Pool

Test 1 Item #	Test 1 p value	Test 2 Item #	Test 2 p value
5	0.16	3	0.20
12	0.85	10	0.87
15	0.75	19	0.74
17	0.69	37	0.63
21	0.23	39	0.25
35	0.35	48	0.30
46	0.65	50	0.69
mean p value =	0.53	mean p value =	0.53

Even if equivalent items cannot be collated, it is still possible to use "p" values to estimate changes in class performance following instruction. Table III shows that seventeen items have been used in a diagnostic pre-test and on the average the class got 67% of the items correct, a little better than the norm of 63%. In the post-test of 20 items, the class still only performed slightly above the norm, 54% compared with 51%, suggesting the instruction had been ineffective.

Table III

Comparing Results from Different Sets of Pool Items

	Mean p value	Mean student performance on test
Test I (17 items)	0.63	0.67
Test II (20 items)	0.51	0.54

Some item pools actually provide pre- and post-instructional "p" values. The usefulness of this information depends very much on the nature of the instruction (how much? how direct? under what kind of conditions?). As it is most unlikely that a given class will replicate the p values (pre-instruction) of the norm sample, the teacher will still have to make an individual judgment about what constitutes reasonable improvement. Is a change from 0.60 (before instruction) to 0.75 (after instruction) as good as an improvement from 0.35 to 0.65? So much depends on how much students have learned from previous instruction and on their ability. Another important matter is the subject under instruction. It is much easier to make progress in subjects like mathematics and a second language where, usually, little is learned outside school than in a subject like English where more is often learned outside the school than inside.

It is vitally important to distinguish teaching to the test objectives, a desirable and eminently sensible practice, and teaching the answers to the test questions, a practice that is as foolish as it is self-defeating.

It is not very sensible to include a difficult "trick" question related to an objective that has not been communicated to learners and towards which there has been no instruction. On the other hand, there is no point in training learners to reproduce memorized codes that will help them answer complex questions and then claiming that the learners have acquired advanced cognitive skills. Teaching to test objectives legitimately implies teaching young people, by any appropriate means, to reproduce the desired and desirable common behavior in response to the relevant stimulus. If analytic behavior is required, then memorized knowledge is not a substitute. But, if analytic behavior is required, it should be taught.

Items from a pool may be used to form a common part of an examination given to all pupils in a particular grade, in a school or in a school district. An estimate of the standards of different classes of schools can be derived from Table IV. In this hypothetical case, three sets of means are shown--from twelfth grade students from five schools on a January examination in English: the mean score on a selection of English pool items related to the common objectives of the school; the mean score on a selection of questions designed and marked independently by each school; and the mean I.Q. score from the schools' records.

Table IV

Results from Twelfth Grade Reading and Literature Tests

	Pool Items		School Items		I.Q. Scores	
	Mean	S.D.	Mean	S.D.	Mean	S.D.
School A, N=212	88.3	23.5	75.3	11.2	108.4	9.7
School B, N=173	82.3	23.2	95.4	10.1	109.6	9.6
School C, N=139	79.8	22.8	69.5	12.1	98.9	13.4
School D, N=257	75.4	23.5	96.1	9.4	102.1	12.8
School E, N=184	62.1	26.4	67.4	13.8	97.5	12.6

41

This is not the place to discuss tests of statistical significance. Suffice it to say that measurement experts can calculate whether differences in means such as those shown in Table IV are statistically significant or whether the probability is that the differences have arisen from chance. For the purposes of this discussion, let us assume that all the differences are statistically significant, and that the probability is less than one in one hundred that any difference has arisen by chance.

At first glance, one might conclude that the results are just common sense. The two schools with the brightest students (Schools A and B) did best on the pool items, and the school with the slowest students (School E) did worst. A closer look suggests some discrepancies. Compare Schools A and B, with very similar mean I.Q. scores. School B performed considerably worse than School A on the pool items but gave much higher scores on its own school items. If both schools shared the objectives of the agreed pool items to the same extent, it would be reasonable to infer that School B has lower standards than has School A. If the marks are to be compared for access to much sought after university programs and scholarships, there are strong grounds for encouraging School A to mark much higher, or School B lower. School D is even more anomalous. It gives the highest school scores and yet has quite mediocre pool scores. The suspicion must be that the school is unfairly rewarding mediocre performance. I must hasten to add that higher standards are not necessarily better than lower standards. After all, one could easily set standards so high that everybody fails. If there are great variations in standards, considerable injustice is done to students who receive lower marks than they would deserve on some universal basis.

Domain-Referenced Tests

A distinction has been drawn between criterion-referenced tests, which are used particularly to determine whether or not learners have achieved certain fairly specific objectives, and norm-referenced tests, which are used to rank learners in terms of their general skills within a given population (pupils of the same age or grade). *Domain-referenced* tests have some characteristics of both. They aim to provide a picture of performance in several different areas.

OISE's Field Development Unit developed a set of tests at the intermediate level, grade seven to nine, in 1978, which were termed domain-referenced tests. In mathematics, the OISE test gives results in the domains of computation, measurement, problem solving, algebra and geometry, and understanding. The tests are also competence based, a principle highly compatible with the use of domains. In each domain and subject, the three levels of competence, honors, and high honors were determined. In addition, achievement norms were developed. Thus, if the norm of achievement in a particular neglected area in a curriculum is seen as being substantively poor, it is possible that even the average student will fail to reach the level of competence.

Many teachers' tests are in essence domain-referenced, in that they are linked to particular domains in the curriculum that the teacher believes have been "covered." The problem with tests such as OISE's Intermediate Tests is that they lose some reliability in the attempt to cover a wide variety of domains. The more varied the course of study, and people's interpretations of it, the less reliable tests are likely to become. They sacrifice some reliability to achieve face validity.

Summary

This chapter has dealt with different kinds of tests, their use and abuse. Tests should be considered in terms of reliability--are the measures reasonably accurate?--and validity--do they measure what the user wants them to measure?

I.Q. tests are useful if the user remembers their limitations. They are only reasonably reliable and valid in the context of a population of homogeneous culture. Even then, the result is not a simple reflection of innate intelligence--the cultural environment of the home is also an influence. Care must be exercised to ensure that expectations are not lowered unfairly as a result of a low performance on an I.Q. test. On the other hand, I.Q. scores are useful: to ensure that highly intelligent children are challenged and that less intelligent children are not accelerated too fast; to ensure that children with various kinds of social, psychological, and neurological disability are not just classified as being "slow;" and to help in the grouping of pupils for instructional purposes. They are also the best way of identifying the intellectually gifted.

Norm-referenced tests are useful for classifying learners in broad categories of achievement level. For example, they help provide a general indication of children's levels in the basic subjects and thus help future guidance. They also provide a universal reference distinct from that of the classroom teacher. A danger is that the objectives tested by norm-referenced tests will be given more precedence by the classroom teacher than other equally important objectives.

Criterion-referenced tests are useful if the teacher wants to assess the performance of learners in a particular, specific task. Domain-referenced tests are useful to provide indications of individual and group performance in areas of achievement, often in terms of defined levels of competence.

Item banks help teachers and administrators evaluate students' achievement of specific objectives in which they are interested. They can also be used to provide a common component in an examination based on a set of common objectives approved by different schools and different teachers. They are flexible tools easily adapted to different teachers' uses.

3 IMPROVING SKILLS IN EVALUATION

Introduction

Although commercial tests and item pools are of great assistance to the classroom teacher, most evaluation is carried out by the teacher, using questions designed for that particular class at that particular time. The good teacher will want to assess learning, provide incentives, classify students and, where appropriate, provide means of certification, as competently as possible. The evaluation will be both formative and summative. The teachers' own assessments will be supplemented by information from commercial tests. Above all, the teacher will want to make sure that individual assessment in the class is as valid and reliable as possible.

Knowledge, Understanding, Thinking

There are various ways of categorizing intellectual goals. Bloom's taxonomies of cognitive and affective objectives are probably the most famous. There is no simple taxonomy that will satisfy all critics and solve all problems. On the other hand, the teacher who avoids taxonomies, arguing that developing classification schemes is like counting the angels that can balance on the point of a needle, is likely to end up with a distorted evaluation scheme. In practice, most tests in schools tend to rely very heavily on straightforward, factual knowledge. Even when questions are intended and appear to be quite subtle, students often succeed best by giving remarkably unsubtle answers. The teacher who asks students to "analyze the causes of the War of 1812" may be subconsciously looking for five points taught in class. Alternatively, after marking five essays by students who are confused between the War of 1812 and the American War of Independence, even the more demanding teacher is likely to think highly of the student who actually gets those five points right. Knowledge questions are not necessarily bad. However, too many tests do not in fact go beyond factual recall.

One fairly simple categorization of objectives (excluding attitudes and values, to which I shall return later) is shown in Table V. Essentially, there are three types of objectives divided into sets of major sub-categories.

Table V

A Simple Taxonomy of Intellectual Objectives

Category	Sub-Category	Evaluation
1.0 Knowledge	1.1 Facts 1.2 Organization 1.3 Display 1.4 Selection	Recall; listing; completion; charts and maps; sequence; formulae; writing; diagrams; graphs
2.0 Understanding	2.1 Comprehension 2.2 Skills	Experiments; recognition of relationships; simple cause and effect; interpretation of figurative and connotative language; interpretation of charts; maps; diagrams; cloze technique; description and paraphrase
3.0 Thinking	3.1 Logic 3.2 Analysis 3.3 Synthesis 3.4 Evaluation	Use of straight and crooked thinking in writing; debate; use of summaries; handling of complex cause and effect; use of independent and dependent variables; decision making; critical judgement.

The third column, evaluation, provides examples of the kind of tasks that are likely to be involved in assessing achievement of the three levels of objective.

The Relative Importance of Objectives

In developing means of assessing students, the teacher must, implicitly or explicitly, first determine the relative importance of different objectives. Obviously, if pupils soon learn that only a few simple factual items will be tested, they will devote their energies principally to learning those few factual items.

The three way categorization of objectives into *must know*, *should know*, and *nice to know* may be helpful. The *must know* items should correspond approximately with the basic pass standard and students should have to be successful in the *nice to know* category if their work is to be classed as excellent.

Let us suppose a middle grade has been studying the Confederation of Canada. The categorization might look something like the following:

Must Know

Date of Confederation; dates and places of conferences leading up to Confederation; provinces entering Confederation in 1867; important provisions of the BNA Act, i.e. major federal and provincial powers, provisions for Protestant instruction in Quebec, provision for a railway to New Brunswick and Nova Scotia; name of the first Prime Minister and his party; residual powers in the hands of the federal government; essential problems of Confederation in the 1981 debate, provisions of the 1982 amendments.

Should Know

Dates and places of Charlottetown, Quebec conferences; subsequent efforts to repatriate the Constitution; changes in the Constitution since 1867; effects of Ontario's and Quebec's domination of Confederation; the French language and the Constitution in 1867 and 1982.

Nice to Know

Some issues in Confederation in relationship with other problems: provincial control of education; French-English problems; dissatisfaction in New Brunswick and Nova Scotia over Confederation in 1867 and in Newfoundland in 1949; effects of the American territorial ambition on Confederation; differences between Canadian and American types of federation; differences between parliamentary democracy and the American constitutional republic; Canada's relationship with Britain in 1867 and today; problems of the Bill of Rights and the Constitution; western alienation; the place of the courts in a parliamentary democracy; freight rates, railways, and Confederation.

The three categories above all include, at least potentially, knowledge, understanding, and thinking. However, it would certainly be possible to phrase entirely factual questions in all three categories. In practice, the *must know* category is likely to be evaluated almost entirely in terms of knowledge.

Let us suppose that a summative test is to be marked out of a score of 100, with a pass mark of 50, honors at 80, and high honors at 90. Approximately 50-60% of the test weight should be based on the *must know* category, approximately 25-20% on the *should know* category, and 15% on the *nice to know* category. That weighting does not imply that half the test questions or half the test time should be devoted to the *must know* category. It may well mean that the more basic questions will have to be weighted disproportionately heavily for the time they will take. This is a good time to emphasize that students should know in advance: the objectives and their relative importance; how they will be tested; and the overall importance of the test and its parts.

Developing Questions to Test Intellectual Skills

Let us assume that teachers are interested in testing knowledge, understanding, and thinking. Certainly, there is neither a single right balance among the three, nor is it possible to distinguish clearly among the three, conceptually or practically. In terms of balance, the first grade teacher giving initial instruction in phonics will be most interested in straightforward knowledge. He will later be interested in applying phonics skills. As far as this particular set of objectives is concerned, he will not be particularly interested in higher level thinking skills. Similarly, a teacher of a first high school course in biology or chemistry is likely to be more interested initially in developing a factual base for the course. On the other hand, a teacher of a 12th grade course in modern Canadian history may determine to give a minimum emphasis to the basic facts, most of which should be already known, and give far greater weight to the application, analysis, and synthesis of the knowledge already assumed.

The three categories are often interwoven. Many facts are virtually worthless without understanding, and almost all understanding and thinking are impossible without a matrix of knowledge. At one time children used to learn to multiply by rote, without having the remotest understanding of what they were doing. That kind of activity is rightly considered of little value today. On the other hand, reaction to that bad practice has left some teachers believing that students can learn to think without facts. One still comes across pitiful examples of secondary students trying to discuss Canada's participation in world affairs without knowing of Canada's involvement in two world wars and the Korean war; the participants and issues in the wars; Canada's membership in NATO; and the meaning of democracy and totalitarianism.

Therefore, although testing in general should involve assessment within all three categories of objectives, all testing in all subjects will not be so balanced. Further, one frequently cannot make a neat allocation of test items among the three categories. It is important to emphasize that it is impossible to create a *teacher proof* item testing higher level intellectual skills. An essay answer to the most obscure and complex question can certainly be taught and memorized, perhaps more easily than answers to multiple choice questions. The essay writer can often skilfully avoid difficult analytical problems.

That raises one of the most difficult questions facing the teacher. How can one teach complex analytical, synthetic, and evaluative competence without teaching sets of specific answers? This is particularly a problem for the teacher of literature, history, and the social sciences. How can one (assuming one should) teach Shakespearean tragedy without encouraging students to memorize examples of symbolism, irony, climax, and tone? Will not students who have memorized such examples, encouraged or not, do better on any test they are likely to receive, even at the university level? If so, why should they not be encouraged to select and memorize such examples? Even in the sciences, recent moves to inductive rather than deductive learning camouflage as much as they solve the problem of teaching the scientific method. Much scientific work is deductive and some people work better deductively than others work inductively. In any event, ideas taught inductively may be learned or applied deductively. At best, most inductive learning in school is simulated, i.e. it recreates the steps of thousands of others. The true scientist sets out alone.

I do not claim to have a complete answer to that most difficult question. The evaluator can always retreat to the position that he is only there to evaluate. I can only counsel sensitivity to the problem. It is important to avoid telling students not to memorize, when in practice knowledge of examples and facts is useful, even necessary. Some will memorize without conscious effort--those who need the conscious effort should not be unfairly penalized. (I speak from the heart here. I was always poor at memorization myself and leaped readily to take the advice of the many teachers who asserted they did not want regurgitation of facts, only to suffer in the examinations that demanded exactly that.) If knowledge is necessary, students should know and they should know which knowledge is of most worth. The teacher should avoid dressing up a fancy question that really only requires an answer memorized from the discussion in class. To revert to my Shakespearean example, it is doubtful how many questions of high intellectual level can be developed on *Macbeth* if it is the only tragedy that has been studied by the 11th grade class. It will be much easier to develop questions demanding analysis, synthesis, evaluation, and critical judgement if learning from *Macbeth* can be applied to one or two other tragedies that have been read and perhaps discussed less fully. Even then, the student who has read and memorized the appropriate Coles' Notes and books of literary criticism on those other tragedies will be at an advantage on most questions. In the land of the blind, the one eyed is king. Comprehensive knowledge is frequently more serviceable than even imaginative ignorance.

Testing Knowledge

The most straightforward, economical, and sensible way to test the facts is simply to ask for them. The obvious exception to this precept is when one wants to test or promote something else at the same time. That exception reasonably leads teachers of history, science, and music to demand, sentence, paragraph, and even essay answers when they are mainly searching for factual recall. They are applying the principle that English is everybody's business.

Straightforward Questions

How many kidneys are there in the human body? (Two)

What is the most important source of energy in Alberta today? (Oil)

What is the symbol for the element lead? (Pb)

The first and last questions presumably have only one possible correct answer, but the teacher must be prepared for an alternative to the second question, e.g. petroleum or oil and gas. That second example is cautionary. There are comparatively few important questions to which there is a single absolutely correct answer, particularly outside the areas of mathematics and science. The fact that one answer has been given in class should never be used as a reason for marking wrong an alternative correct answer. A frequently made mistake is to have students *fill in the blanks*, with only one answer being accepted for each blank, and yet where the incomplete sentences are ambiguous. *Filling in blanks* is best used in either one of the following circumstances:

1. The blank can be possibly filled by only one, single, factual answer;
2. Multiple choices are provided for the blanks and there is a single right or best answer for each blank.

The reader may argue that *filling in blanks* is perfectly appropriate in broader situations than the above provided the teacher is prepared to be flexible in marking. My guess is that that proviso is rarely satisfied. The advantage of *filling in blanks* is precisely that the questions are easily scored. Imaginative, divergent thinking can easily suffer. Far too often teachers fall back on the defence for their answer that, "This is what I said in class. You should have been listening."

True and False Questions

By definition, these questions can only be used for straightforward factual matters. This type of question is generally and appropriately out of favor. It has often been abused through extension to non-factual matters. For matters of fact, there are nearly always more efficient and effective ways of questioning -- straightforward informational questions, filling in blanks, etc. If true/false questions are used, the number of wrong answers must be subtracted from the number of right answers. Otherwise the results are not reliable. If the questions are factual and unambiguous, and if wrongs are subtracted from rights, true/false questions can be used reliably in the physical and social sciences.

Teachers of English often insist that questions of factual comprehension be answered in complete sentences. This is a good practice as it encourages correct writing. However, skilful writers can often use that skill to obscure rather than reveal ignorance. The essay question is therefore generally a not very good measure of knowledge unless extremely precise instructions are given. Further, the marking key must be tied closely to the instructions given to answer the question.

Poor Example: Twelfth Grade Geography
"Describe the climate of the Cape Province of South Africa and its implication for agriculture."

Better Example:
"Describe the climate of the Cape Province of South Africa, including precipitation and its incidence, and prevailing temperatures in the various seasons. Show to what extent the agriculture in the Province is adapted to the climatic conditions and discuss the possibilities for other types of agricultural production."

Time: 30 min.; Weight: 25%, of which: 10-Climate
7-Agriculture
3-Other Possibilities
5-Expression & Organization

If written answers are not marked by the teacher, their effectiveness in encouraging writing is unlikely to be very great. If they are marked, the students must know how much weight is given to writing skills.

Multiple Choice Questions

Contrary to popular opinion, multiple choice questions are not the most useful for purely factual matters. They are best for testing general understanding and the ability to analyze information. Indeed, they do have an important disadvantage--that one can frequently obtain the right factual answer by eliminating wrong answers. In other words, on strictly factual matters, it is perfectly possible for students to get right answers without having complete recall of all the facts. That problem could be serious, as in the following questions:

Which of the following foods is a good source of protein?

A. cheese*
B. candy bars
C. potato chips
D. French fries

There are many reasons why students might pick the right answer without having the slightest idea about which foods contain protein.

A better form of questions would be the following:

Which of the following foods is the best source of protein?

A. French fries
B. hamburger*
C. apples
D. granola bars

The second question does not give cues; indeed, students may be tempted to pick apples or granola bars because they believe these are approved foods.

Even when the multiple choice question is not particularly badly constructed, elimination can prove a problem. For example, many students might get the following question without knowing that Charlottetown is the capital of Prince Edward Island:

Which of the following is the capital of Prince Edward Island?

A. St. John's

B. Vancouver

C. Moncton

D. Charlottetown

A little knowledge eliminates St. John's and Vancouver. One may remember that Moncton is not in one's list of capitals. Charlottetown must be it then! Far better to test knowledge of the name together with knowledge of its position.

In general, teachers have comparatively little problem testing purely factual matters. The most frequent problems they do encounter are:

(1) demanding a single right factual answer when there is more than one possible answer;

(2) asking rather pretentious questions that really reduce to a factual answer;

(3) not ensuring that students know what knowledge they are responsible for;

(4) failing to draw reasonable dividing lines between important and trivial or abstruse knowledge.

Alternative Approaches to Testing Knowledge

Much knowledge is of little value outside the context in which it is likely to be used. It is therefore desirable to make sure that students not only possess knowledge, but can apply it. Very often that means moving from knowledge to understanding. Before turning to that problem, there are circumstances where the application of knowledge can be tested at a fairly rudimentary level.

In geography, knowledge of places, rivers, and mountains should be applied to maps and globes (and that application scarcely justifies the term *understanding*). In science it is useful to make sure learners can draw a diagram and label the parts appropriately. In mathematics, children should be able to compute and measure in real situations, and not just from abstract examples.[1] In spelling, children should not only be able to spell words and recite memorized definitions, but should also be able to use them intelligently.

---

[1] In saying this, I am not espousing the false idea that children can only learn by doing. I am merely pointing out that if teachers want practical application, that application should, where possible, be practised.

Examples: Geography

On the map of Canada: Name 1. line of latitude A
2. capital city B
3. lake C
4. the mineral mining region D
5. major economic production in region E

Spelling

The teacher provides the oral instruction:

"Use the following words in sentences that show you understand each word's meaning. *Gratitude, gratitude*
Resplendent, resplendent"

The following would be wrong answers:
I felt a great deal of *gratitude* when I saw him.
She looked *resplendent* when she came into the room.

The following would be right answers:
I felt a great deal of *gratitude* for all the good things he had done for me.
She looked *resplendent* in her fine new gown and beautiful silk slippers as she arrived at the ball.

Which of the following words has the same *root* as *gratitude?:* grate, grateful,* grater, grating.

Science

Instead of learning various types of classifications pupils can be asked to apply the classification systems. For example, trees can be identified in various ways, e.g. presence of terminal buds, opposite or alternate buds, leaves serrated or smooth.

These examples lead one to more complex categories of knowledge and even to understanding. For example, I question the usefulness of knowing that

Charlottetown is the capital of Prince Edward Island if one does not know where the island is on the map of Canada. There is, I believe, no value in knowing how to spell a word one cannot use. Obviously, the precise dividing line between knowledge and understanding is arbitrary and judgemental. I am making a distinction for the practical purpose of illustrating ways of testing different levels and kinds of achievement.

Matching questions are particularly useful in quickly checking knowledge in formative evaluation. The chemistry teacher can ask students to match elements with named compounds of which they form a part. The geography teacher can ask for a link to be made between climatic descriptions and places. The matching test is fairly easy to construct and can even be easily marked by the pupils themselves.

Testing Understanding and Thinking

The two best traditional ways of assessing more complex intellectual objectives are the use of multiple choice and paragraph or essay type questions. Before looking at them more closely, it is worth considering a few other options.

Oral Presentations

A later section deals with the evaluation of the achievement of expressive objectives. In this context, the relevant question is: "Are oral presentations good ways of evaluating intellectual content?" If they are to be considered good, they must be both valid and reliable.

Oral evaluation is a controversial topic. If listening skills and oral expression are important, should they not form an important part of an evaluation scheme? I would be very cautious about embracing oral evaluation too warmly. Certainly, I am not denying that listening and speaking skills should be evaluated systematically to the extent that they form distinct objectives in language and social science courses. But I am asserting that one should be very careful in using oral work regularly as a means of assessing general academic achievement.

Even competence in writing interferes seriously with the evaluation of intellectual skills in, say, history. That is one reason why objective, multiple choice questions are very important supplementary tools. The problems of interference by speaking and writing abilities are analogous. Nevertheless, it can be argued that ability in written expression is almost inextricable from the skill of the historian; the same case is less strong for oral expression.[1] Let me explain what I mean here. Differences in ability in oral expression interfere much more greatly than do differences in writing ability with an evaluation of the intellectual content of what is said. A student may be shy, lacking self-confidence, may have heavily accented English, may have a tendency to stutter, may have a poor speaking voice, and may have an unprepossessing appearance. Those qualities make the individual a poor public speaker, but they do not constitute grounds for the provision of a low assessment in the understanding of history or geography.

It is true that the ability to respond to questions raised about statements made by a speaker is evidence of understanding and, probably, of thinking. Similarly, the depth of questions raised on a topic is an indication of the kind of understanding and thinking that has been achieved. One way to measure a student's ability to defend a thesis is very obviously to have the student present it and answer questions on it. However obvious that may be, reliability remains a problem. Essay marks tend to be unreliable. As anyone who has judged competitive public speaking knows, evaluation of oral presentations tends to be even more unreliable. Here are two, very tentative, suggestions of ways to reduce that unreliability.

First, have students share the assessment. Five different students are selected to score, independently, every speaker. Friends and enemies should be avoided. One might give the presenter, privately, six possible evaluators and permit one veto. The scoring team represents both sexes and various levels of ability. On each criterion the median rather than the average score should be selected for comparison with that of the teacher. Use of the median, i.e. the mid score, avoids the problem of powerful bias or extreme judgements. Second, use very specific criteria, particularly if one wants to concentrate on substantive issues and reduce reliance on expressive abilities.

---

[1] I am not strongly supportive of the argument that written expression is inextricably linked with skills in a discipline. It rests on the assumption that the major purpose of teaching history is the production of historians, a rather fanciful assumption at the seventh grade level. The point I am making is that the case is even weaker for oral expression.

Table VI provides one set of criteria. The proportion of marks I have devoted to answering questions probably seems low. The problem is one of reliability. If questions come from different students they are likely to vary greatly in terms of difficulty, making judgement of the response very hard. Some presentations may elicit very few questions, making marking extremely difficult.

If the class appears to be working seriously, the teacher may decide to combine the median and teacher's score, as shown in Table VI. The entire score sheet should be provided to the student concerned, but the scores should probably not be identified. The completed score sheet should not be made available to the class as a whole. With the approval of a volunteer guinea pig, a training session is useful to improve scoring, and indirectly to improve performance. In a training session all scores, including the teacher's, can be made public on the chalkboard or overhead projector.

The disadvantage of the above procedure, apart from its possible unreliability, is the amount of time it consumes. Evaluation of a class through oral expression will take at least fifteen instructional hours. A combined essay and objective test could provide at least as reliable results in one hour. Obviously there are some complementary instructional advantages in the oral test. It should certainly be used, but sparingly.

Cloze Tests

The Cloze test is essentially a test of comprehension, although it can also be adapted to test knowledge of language. In its simplest form it is extremely easy to develop. One may, for example, select a passage whose comprehension one wants to test, and eliminate every tenth word. The students are simply asked to fill in the gaps.

Cloze methods are obviously used most to test comprehension in both first and second languages. Cloze tests can be put in multiple choice format, where there may be four or more options for each gap.

Table VI

Criteria for the Evaluation of a Social Sciences Presentation

		Scorers:							
Criteria	Possible Points	1	2	3	4	5	Median	Teacher	Final
Presentation of Basic Thesis	15	7	7	6	4	8	7	8	7.5
Logical Support of Thesis	15	8	8	4	3	9	8	7	7.5
Evidence of Research	10	9	8	6	5	9	8	9	8.5
Relevance of Evidence	10	9	9	7	6	8	8	7	7.5
Consideration of Counter Evidence	10	6	6	6	1	8	6	4	5.0
Reasonableness of Conclusion	10	9	9	6	3	10	9	7	8.0
Clarity of Conclusion	10	10	10	8	6	10	10	10	10.0
Method of Presentation	10	8	9	7	5	8	8	8	8.0
Competence in Dealing with Questions	10	8	8	2	3	8	8	6	7.0
	TOTAL 100								69.0

This method is useful in testing a second language. Example:

(1) (Le, La, Les, Des) chat est sur (2) (le, la, les, des) table dans la (3) (cuisine, lit, auto, jardin). (4) (Il, Elle, Ils, Elles) s'appelle Marc.

In other subjects, composition of good, valid Cloze tests is not easy. Whereas in French it is often useful to test or review a number of points of grammar, vocabulary, and syntax, one's objectives in other subjects are likely to be less easily collated in a single passage. It is extremely easy to revert to testing knowledge almost exclusively. The omission of multiple choices will permit the exercise of more imagination, but scoring will be correspondingly more difficult and therefore less reliable. For example, how does one compare an imaginative but not quite correct answer with a simple, factual one? Is it fair to give extra credit for imagination if imaginative answers have not been requested? If imaginative answers are requested, where does one draw the line? And why?

The Cloze test, unsurprisingly, has been sorely abused, particularly because it has two apparent advantages over a regular multiple choice test --it is much more easily constructed and has the glamor of novelty.

Matching Questions

Good matching questions can test understanding as well as knowledge. In general, however, it is probably more straightforward and reliable to use multiple choice questions for that purpose. Complex matching questions can easily be confusing. There is no point in misleading or tricking people into wrong answers. The following is an example of a matching question that tests some understanding as well as knowledge of economics. The question is perfectly acceptable but a multiple choice format could also cover the same ground. The multiple choice format might also give the teacher a clearer picture of just where the problems lie.

Sample Question:

Make the best possible fit between the terms in Column A and the conditions described in Column B:

Column A

1. inflation
2. deflation
3. unemployment
4. underemployment
5. overpopulation
6. recession
7. elasticity

Column B

a. "Wages, salaries and prices in Israel are rising very much faster than production."

b. "The gross national product, expressed in constant dollars, declined in Canada during the last quarter of 1980."

c. "Recurrent droughts in the savanna lands of southern Africa make it unlikely that the country can ever be self-sufficient."

d. "Many jobs carried out by a single person or a machine in western countries require a bureaucratic hierarchy for accomplishment in India."

e. "In Britain during 1979 the available jobs declined while the available workers increased."

f. "The demand for winter vacations in Florida appears to vary greatly in Canada depending on the status of the economy in general and the value of the dollar in particular."

g. "Monitorists, such as Milton Friedman, advise strict control of the money supply as a response to overheated economies."

Answers: 1. a 2. g 3. e 4. d 5. c 6. b 7. f

Note that the instruction uses the expression "Best possible fit." The word "best" may be seen as a weasel word but it is in fact a necessary qualifier. If the instruction just said, "Match Column ....." a student may claim that "overheated economies" clearly indicated inflation not deflation. However, the remedy in condition "g" certainly indicates "deflation"; and "inflation" is the only reasonable answer for condition "a." Questions of understanding often require judgements and good tests that discriminate among the most able will require fine distinctions to be made.

Multiple Choice Questions

The multiple choice question is the preferred option for substantial parts of summative evaluations in a large number of subject areas in the intermediate and senior levels. Those subject areas include: history, geography, economics, literature, physics, biology, chemistry, home economics, electricity, electronics, and health. To a lesser extent, it is useful in music, art, languages, mathematics, and most technical subjects. I am not suggesting that there should be total reliance on multiple choice questions in summative evaluation in any of the above subjects.

Let us examine some of the criticisms of the multiple choice test:

1. It discourages the use of written English and helps to produce a generation of illiterates.

Heavy reliance on the multiple choice question will undoubtedly have this effect. There was widespread feeling in the U.S.A. at one time that heavy emphasis on multiple choice college entrance board tests led to a de-emphasis of written English in the nation's high schools. I would certainly recommend that every student in school from the age of 8 up be examined specifically in written English skills and that students' competence in this area be an important component of scholastic recognition. I would also recommend that all major tests and examinations in the social sciences and humanities include a component demanding written expression. Even in the sciences, the ability to write scientific reports should be regularly evaluated.

2. It trivializes the curriculum by reducing it to a few learnable facts.

It is simply wrong to say that the multiple choice question only tests knowledge. Learning ability tests, for example, are usually in a multiple choice format. They are all, to a greater or lesser extent, divorced from knowledge. An abundance of factual knowledge, of itself, will not produce a high I.Q. score. Even so, it is true that the multiple choice question cannot readily be used to test the more complex intellectual skills of synthesis and evaluation. The multiple choice question is unlikely to favor creativity. One must be careful to avoid the crooked argument that because some multiple choice tests are juvenile, faulty, invalid, and unreliable, all multiple choice tests must be poor. The cross the multiple choice test has to bear is visibility. Everyone sees the teacher's mistakes in a multiple choice test. The teacher marking an essay may miss most of the points made by the student and be totally ignorant of the others, but the mark reveals nothing.

3. It mainly tests intelligence and reading ability.

Results from good multiple choice tests will probably correlate more closely with I.Q. scores than will results from essay tests, particularly if the essay is written as part of an out of school assignment. Analytical intelligence, for example, will contribute to scores on multiple choice questions testing analytical skills. Presumably, schools are interested in promoting both analytical intelligence and the conscientiousness, conformity, and perseverance that are more strongly reflected in home assignments.

Certainly, multiple choice questions depend on reading ability, just as essay answers depend on writing ability. That does not seem unreasonable. The student who is to progress in the social and physical sciences should be able to read and write with proficiency. Even so, a good multiple choice test in chemistry cannot possibly be answered solely on the basis of intelligence and reading ability.

4. It is extremely difficult to develop.

Good multiple choice questions are indeed very difficult to develop. In partial compensation, they are very easy to mark. Furthermore, the items can be kept in pools and used on various occasions in different ways. Mention has already been made of Ontario's item pools in a number of subject areas. School districts might also consider developing their own pools to match objectives they believe to be different from provincial objectives.

Set against those criticisms, the multiple choice test has the following advantages:

1. Multiple choice tests can be extremely reliable.

The questions normally have right or wrong answers. Anyone marking them will get the same results. If the questions are phrased carefully, the test is likely to have very high reliability. That cannot be said of most oral answers or answers written in essay or paragraph form.

2. Face validity is fairly easy to establish.

It is possible to see just what the question is testing. That is often not the case with a written answer. Written answers often obscure ignorance as much as they clarify knowledge. Multiple choice questions are excellent vehicles for valid tests of: knowledge and its applications; general and specific comprehension; inference; analytical and logical thinking; and understanding of relationships and experimental effects. They can also be used to test the recognition of effective synthesis and evaluation although not the acts of synthesizing and evaluation.

3. They are easy and economical to use.

Multiple choice questions often take less than one minute for the student to answer. Quite reliable tests can be administered to large groups in less than 30 minutes. Pools of questions can be drawn upon over a period of years.

Developing Multiple Choice Questions

Multiple choice questions consist of two parts, a stem and a set of options, one of which is correct, the remainder being distractors. The number of options varies from three to five. The use of only three options encourages guessing, and it is therefore desirable to subtract half the wrong answers from the right answers to arrive at the total score. The use of five options tends to stretch the test maker's imagination and lead to useless or even facetious distractors. Little reliability is lost by the use of four rather than five options, but four are generally preferable to three.

Considerable experimentation has been carried out with different approaches to the multiple choice question. One problem of the multiple choice question, when it is used to make quite fine distinctions at a high intellectual level, is that individual judgements will inevitably differ. A good question, by definition, will generally be answered correctly by experts. But when fine distinctions are called for, the occasional expert will sometimes get the question "wrong." Students understandably complain that "even the teacher got that one wrong." One approach to that problem has been differential scoring--with three marks for the right answer, two for the second best, one for the third best and zero for the worst. Unfortunately, it is difficult enough to develop sophisticated tests with a right answer and three distractors. It has proved impossible to produce such fine gradations as partly right answers with any reliability. The common four option (one correct answer and three distractors) variant of the multiple choice test proves generally to be the best bet.

The following suggestions are useful to bear in mind when developing multiple choice questions. Most of them are not absolute rules. There are sometimes reasons for ignoring the so-called rules in question development. They should be used as guides for reference.

Suggestions for Developing Good Multiple Choice Questions

1. The stem and option should be complete and make sense, e.g.

   Good

   Which of the following is a common and dangerous practice of bicycle riders?

   A. Failing to halt at stop signs.*
   B. Taking one hand off the handle bar to wave to a friend.
   C. Forgetting to bring one's puncture repair outfit.
   D. Failing to dismount on a steep hill.

   Bad

   Which of the following is a common and dangerous practice of bicycle riders?

   A. Riding to school on a bike.
   B. Knocking down pedestrians.
   C. People who go out on icy roads.*
   D. Signalling.

2. All options should be essentially parallel in terms of grammar, syntax, length, and type of answer, e.g.

   Good

   Why did John Dodds want to kill his father?

   A. He was angry that his father's will left everything to John's sister.
   B. He was jealous of his sister who was attractive and successful.
   C. He wanted to spare his mother from his father's foul temper.
   D. He thought that this was his only chance to get rich.

   Bad

   Why did John Dodds want to kill his father?

   A. Anger.
   B. He was jealous because of the lack of attention given to him compared with that given to his sister.

C. Because his mother was treated so badly by his father.

D. The only way to get his own back.

3. Distractors should not be correct in different ways, with one way wanted by the teacher, e.g.

Good

Which of the following sentences most clearly describes the function of the arteries in the human body?

A. They allow the blood to combine with oxygen to provide food.

B. They connect the lungs and other vital organs with the heart.

C. They allow the blood to flow from the heart to various parts of the body.*

D. They allow the blood to flow from the body's extremities to the heart.

Note: C is clearly the best answer - note the use of "most clearly" in the stem. A danger in this question is the use of opposite answers in C and D. If this happens very often it may become a cue that one of the two opposites is correct.

Bad

Why did Britain declare war on Germany in 1939?

A. Germany had occupied the Polish Corridor.*

B. Germany was continuing a policy of territorial expansion.

C. The Soviet Union had occupied Poland.

D. Britain realized that a policy of appeasement would not work.

Note: The teacher wants the direct answer A. However, it is not reasonable to reject answers B and D which, in some ways, are better answers.

4. Only one point should be tested by a single question, e.g.

Good

A coat discounted by 8% is priced at \$66.24. What was the original price?

A. \$71.54

B. \$5.30

C. \$71.00

D. \$72.00*

Note: Make sure the distractors include common errors and an answer approximating the correct one, unless one is testing estimation.

Less Good

A train travelling at a speed of 120 km/h leaves Kingston at 07:05. What time will it arrive at Oshawa, 201.3 km further east?

A. 08:05

B. 08:45

C. 08:46*

D. 09:45

Note: This is not a bad question but some pupils will get the tricky calculation wrong even if they understand the method. If a test of division is required it is probably better to test it directly.

This rule should not be overinterpreted. It is sometimes used as an argument for only asking very simple questions. Application of a formula in physics presupposes computational skills or the ability to use a calculator. If computational skills are necessary in physics, then it is certainly not unreasonable to demand competence in their use. That does not mean students in physics should spend most of their time during a test working out incredibly complex multiplication and division questions.

Similarly, it is not unreasonable to demand in a single question everyday basic knowledge together with some advanced complex thinking. It is unwise to demand abstruse knowledge and advanced abstract thinking in the same question.

5. The right answer should not stand out for the wrong reason.

Reference has already been made to syntax and length, which may give clues to the right answer. Elaborate, sophisticated language is also sometimes an inadvertent clue. One sees distractors that are just plain silly placed next to a rather sophisticated, correct answer. In general, distractors should make some kind of sense. It should not be possible for an intelligent person to get most questions right without having mastered the objectives of the questions as opposed to the testing. For example, it would not require much knowledge of Buddhism to answer the following question correctly:

Which of the following is an important component of Buddhism?

A. One should free the mind from materialistic temptations.*

B. You can do whatever you like as long as it makes you feel good.

C. You should always follow Mohammed's teachings in everything.

D. It is all right to kill your enemies as long as they are not Buddhists.

Option A. stands out as being different in language and structure. The distractors all appear highly improbable.

6. Trick questions are unlikely to test the stated objectives.

An exception to this rule is the problem solving question where the trick may be part of the objective being tested. If the teacher "catches out" most of the class with a trick question, the question only serves to test mastery of the trick not whatever is left in the question.

7. Answers to earlier questions should not provide clues to later questions.

This rule too must be treated with care. Generally speaking, if one changes the order of questions in an attempt to conceal a relationship between two questions, one has a problem. The students should not be able to improve their scores just because they happen to remember that the stem of one question answers another later question. But that does not mean that one should eliminate a question that tests a very similar point to that tested by another question in a slightly different context. On the contrary, that is often a very useful device, provided the point being tested is worth two questions. The two questions provide validity checks for each other. The reductio ad absurdum would be to test only one point very reliably--one would then have a criterion referenced test.

8. Correct English should be used and spelling should be consistent, unless bad English or incorrect spelling is being tested.

Nothing is more irritating than a "correct" answer written in poor English with incorrect spelling. How can the student be expected to identify such an answer as correct? None of us is perfect. Multiple choice tests should be checked at least for their language by a colleague before being used.

The question of whether multiple choice questions should be used to test spelling and English usage is moot. Some teachers believe that bad examples merely serve to reinforce false notions the students may already have. (Indeed, many English teachers abhor all multiple choice language tests. They argue that language should be marked as a part of the student's normal writing. Unfortunately, their system is usually unreliable and students who use very simple English correctly "beat the system.")

A cumbersome alternative to the multiple choice question is directed sentence construction, e.g. "Write a sentence using the adjective *different* followed by a prepositional phrase." Many teachers also object to that approach on the grounds that it involves teaching and naming grammatical structures.

9. "None of the above" and "All of the above" options should be used sparingly if at all.

If they are used, then they must be sometimes right and sometimes wrong. The use of "None of the above" limits the depth and subtlety of the questions that can be used. Obviously, distractors must be clearly right or wrong when used in conjunction with "None of the above." Similarly, with "All of the above" there have to be very strong distinctions between right answers and distractors. If "All of the above" is correct, then the three distractors must be equally attractive, e.g.

Acceptable

Pierre Trudeau's most significant accomplishment as Prime Minister has been:

A. Introducing a Canadian Flag.
B. Taking Canada into the Korean War.
C. Taking Canada out of the North Atlantic Treaty Organization.
D. None of the above.*

Note: There is always the danger that the ignorant student who, as well as being ignorant, does not think that Pierre Trudeau has had any significant accomplishment will pick "D" for the wrong reason.

Bad

Lester Pearson's most significant contribution as Prime Minister was:

A. Integrating Canada more closely with the U.S.A.
B. Bringing about a closer relationship between Canada and Britain.
C. Bringing about good relations between Quebec and the rest of Canada.

D. None of the above.*

Note: A. is a poor distractor. Some would consider it accurate in terms of Pearson's contribution, although they might not consider it valuable. Does "significant" imply valuable? C. is also poor as die hard Liberals might argue that Pearson did help in this respect, despite the FLQ.

10. Negative questions should be used sparingly.

Some test constructors argue that negative questions are illegitimate. Yet, test statistics sometimes show negative items to be performing and discriminating in an excellent manner. Negative questions can quite easily be confusing. They are therefore not very appropriate for the young child, for people whose first language is not English, and for those whose intellectual reading skills are weak. In contrast if one is to discriminate among good readers with fairly high intellectual skills, a few negative questions may be extremely useful, e.g.

Good

Dr. Whalen: "I don't understand why Mary Inglis should so ignorantly attack my research findings. Mrs. Inglis is not famous for research in animal genetics. I discovered that black squirrels in my Toronto sample were the only ones carrying the mutation which we believe leads to cancerous abnormalities. Not a single grey squirrel had the mutation. There is no reason for her to ridicule my suggestion that we eliminate all black squirrels from the population."

Which of the following is not a *logical* response to Dr. Whalen's comments?

A. Mrs. Inglis may be a woman but she may know just as much about squirrels as Dr. Whalen.*

B. We need to know much more about the problem and the proportion of black squirrels carrying the mutation before making recommendations such as Dr. Whalen's.

C. We should be less interested in Mrs. Inglis' background than in the substance of her argument.

D. We should be looking well beyond the Toronto sample before coming to decisions to get rid of black squirrels.

Note: Option "A" does not deal with any point made by Dr. Whalen.

Poor

Which of the following are *not* antonyms?

A. uninterrupted ... continuous.*

B. expire ... survive.

C. inflammable ... non-flammable.

D. felicitous ... gauche.

Note: Although hardly unacceptable, the above item could cause confusion with its triple negative (not, antonym, uninterrupted).

Using the Multiple Choice Item to Test High Level Objectives

The higher the level of objective, the more difficult it is to develop a good question. One must accept that on most complex items, the occasional good student, and expert, will get a "wrong" answer. Occasionally, extreme knowledge, unusual understanding and divergent thinking will lead a very good thinker to a less obvious interpretation. This is unfortunate but inevitable if fine distinctions are to be made. This problem is not in any way confined to the multiple choice question but the difficulty is certainly made explicit in the multiple choice format. The divergent thought in an essay can easily be seen as a sign of ignorance by a teacher marking against the clock or by a teacher who is simply less knowledgeable than the student in the particular area. In the higher grade levels, most teachers simply cannot be expected to be expert in all the conceivable areas in which a good student may write. At least, problems in multiple choice questions can be checked and queried. The problem (of good students getting "wrong" answers) becomes very serious when bright students generally do no better than poor students on a particular item. I shall deal with that problem later.

The Ontario Assessment Instrument Pool in reading at the intermediate level is concentrated particularly on higher level thinking objectives. There is not space available here to include lengthy written excerpts and the associated multiple choice items from that pool. One approach used was to simulate classroom debates on controversial topics. The questions examined ways of challenging the logic of the various participants in the classroom discussion. Other items demanded that the respondent make appropriate comparisons and analogies between different prose and verse selections. Good high level items in the arts and social sciences usually require fairly lengthy prose, failing that, a very clearly articulated program of studies. The following items are taken from the Intermediate Tests. These domain-referenced tests were designed for students at the eighth grade level. They do not generally reflect very high level objectives but the selected questions do go beyond the usual recall and comprehension.

Canadian Studies

Which of the following statements best describes the way in which democratic countries (such as Canada) differ from communist and totalitarian countries (such as Russia and Chile)?

A. Democratic governments only pass the laws that most of the people want.

B. Democratic governments take into account the best interests of the people before they pass laws.

C. Democratic governments ask people what they want before they pass laws.

D. In democratic countries people are always free to do what they want, even if the government does not like it.

E. Democratic governments are freely elected and it is possible to change the government at regular intervals.*

Note: This question demands some comprehension and some analysis. The use of the word "always" may be seen as a cue--very few things are either "always" or "never" true. The right answer gets at the essence of democracy.

Geography

The following table gives some climatic data for Vancouver, Toronto, Winnipeg, Halifax, and Edmonton. Which one is correct for Toronto?

	Mean July Temperature	Mean January Temperature	Mean Annual Precipitation
	^{0}C	^{0}C	mm
A.	18	-4	1300
*B.	22	-4	740
C.	20	-17	470
D.	17	-13	420
E.	18	3	1360

Note: This question demands knowledge, analysis, and the application of relationships.

Health

Which of the following would be a *poor* daily diet for a teenager?

*A. Breakfast: coffee, toast
Lunch: French fries, soft drink
Supper: spaghetti and tomato sauce, cake, milk, ice-cream, fruit

B. Breakfast: cereal, milk, toast, juice
Lunch: peanut butter, bread, juice
Supper: fish, potatoes, peas, ice-cream, fruit

C. Breakfast: bacon and egg, juice, milk
Lunch: hamburger, French fries, salad, juice
Supper: pork, potatoes, peas, ice-cream, fruit

D. Breakfast: cereal, milk, juice, toast
Lunch: tunafish sandwiches, orange, milk
Supper: omelette, potatoes, salad, ice-cream, fruit

E. Breakfast: tea, juice, boiled egg, toast
Lunch: hot-dog, raw carrots, chocolate milk, cookies
Supper: fish, potatoes, green beans, ice-cream, cheese

Note: This question demands application of a reasonable level of comprehension.

Science

Mr. Nucci asked the grade 8 science class to suggest *hypotheses* as the class was going to carry out some scientific investigations.
Which of the following suggestions is a hypothesis?

A. Both aluminum and tin are elements.
B. Circles are bigger than squares.
C. Liquids are more fluid than solids.
D. Plants grow more rapidly in a south window than one facing north.*
E. The condition of rust represents a chemical change.

Note: This question demands analytical application of understanding of the concept. The question is marred by the length of the correct answer.

Synthesis and evaluation are two important objectives generally best evaluated by paragraph or essay questions. However, it is possible to see if students can at least recognize good synthesis when they see it. In the Ontario reading pool, there are two opposing statements on the problem of sexism in the secondary school. One of the items accompanying the passages asks students to pick the best synthesis of the two arguments. Similarly, different evaluations, critiques, can be compared in terms of their technical qualities. A sound evaluation, e.g. a critique of a film, should have two basic components--an appraisal of the worthwhileness of the enterprise itself and an appraisal of the enterprise's success in achieving its goals. A multiple choice question testing the inclusion of those two components at least tests student's abilities to apply the principles of evaluation. They can only prove they can do it by performing an evaluation, orally or in writing.

Making Sure Multiple Choice Questions Are Working

One of the advantages of a multiple choice test, or section of a larger test, is that it is fairly easy to check how the various questions are working. There are various formulae for carrying out detailed checks of reliability, usually requiring access to a computer. However, it is fairly easy for the classroom teacher to make informal checks.

The simplest way to check the question is to discuss the answers with the class. If the questions are being used for the first time, they are probably not all perfect, particularly if they are testing complex intellectual skills. Students may be given the opportunity to argue for their chosen distractor--and even given partial credit on the basis of their argument. Obviously, everyone else with the same answer must also be given that credit. Where there is dispute (or where the teacher has doubts) it is useful to ask students with the correct (according to the marking scheme) answer to give their opinion. If a much higher proportion of more successful than less successful students (judged in terms of the results on the whole test) has the question right, it is more likely to be a fair one. The teacher should also consider excluding questions completely if they turn out to have been hopelessly ambiguous. If good students are just as often wrong on a particular question as poor students, it is likely to be a poor question and a strong candidate for exclusion. The fact that most of the class picks distractor "C" does not mean that the question is necessarily a poor one. It may just be that distractor "C" is working as a very good distractor; option "B" may still be clearly correct. The majority is often wrong.

The same kind of check can be carried out a little more formally using a chart like that in Table VII. The table shows how a teacher may compare the performance of the best students over all with that of the worst students over all. The assumption is that a good question will discriminate between better and worse students. Question #9 is not a poor question just because nearly everyone gets it right. It may represent the basic competence, *must know* category, that the teacher really wants. Question #15, is, one hopes, in the *nice to know* category. In that sense it is a good question as it discriminates rather well. Question #18 is a much more doubtful case. The distribution of right answers is even throughout the class. The question is intended to be a test of high level skills, and probably is not working. I say "probably" because interpretation must take into account

Table VII

Checking the Usefulness of Multiple Choice Questions

Questions For Consideration	Top 10 Students		Bottom 10 Students		All Students N=35	
	Correct		Correct		Correct	
	#	%	#	%	#	%
# 9	10	100	10	100	33	94
# 15	4	40	0	0	7	20
# 18	6	60	7	70	22	63
# 27	2	20	3	30	5	14

the specific situation. If the class were homogeneously highly intelligent, that statistic could have resulted from an excellent question on a topic not taught. A more likely possibility is that the question is one of detailed and irrelevant fact--forgotten equally by good and bad students. Yet another possibility is that the question is based on an important but not taught fact. Finally, the question may be ambiguous. Of the four questions, #27 is the strongest candidate for rejection. Very few students are able to tackle it successfully and the poor students have actually done better than the good students. Possible explanations are that the students have just been guessing and that the question is ambiguous. "If the statistics in the table do not provide a clear diagnosis, why bother to collate them?" one may ask. The answer is that questions that do not appear to work should be examined very carefully. If a good reason for their failure is not found, they should be discarded. A check of this kind can also be combined with review to help students (and the teacher) diagnose areas of difficulty. A reliability check of some kind should be carried out before items are carried over into a bank for future use.

Correction for Guessing

If four or five options are used, it is generally not necessary to correct for guessing, provided that the test is not unusually difficult. If no correction for guessing is to be made, students should be told to make intelligent guesses once they have eliminated some distractors. But they should not waste too much time on questions they cannot do.

If a correction is to be made for guessing, students should be so informed. If there are five options, one quarter of the wrong answers should be deducted. If there are four options, one third of the wrong answers should be deducted. The disadvantage of making corrections for guessing is that the process affects students differently. It may make some quite good students overly cautious and thus omit items requiring fine and difficult judgement. The following situations make corrections for guessing more desirable:

1. The test is largely based on knowledge;
2. The answers are usually clearly right or wrong;
3. There is a tendency for pupils to guess frivolously.

The Paragraph and Essay Question

The assessment of written English is an enormous study in itself, to which I cannot possibly do justice for two reasons: lack of space and lack of competence. One slightly comforting thought for the ignorant reader and the ignorant author is that very little is clearly established from all the research in this area. No one knows for sure the best ways either to evaluate or to teach written English. The following remarks then are more than usually tentative and subjective.

Assessing English Qua English - Holistic Scoring

One great difficulty with the evaluation of written English is the problem of reliability. Many experiments have shown that different markers vary greatly in their assessment of written compositions. The most reliable way to evaluate written compositions is by the use of holistic scoring. This method involves two or more teachers independently giving an impression score, on a scale of, for example, 1 to 10 for every essay. Under ideal conditions, where all teachers can be calibrated on the basis of their scoring of a common set of essays, reliability can be close to that of a good multiple choice test. That is not possible in the normal school situation. Even so, if teachers, through discussion and practice, can ensure that their standards, not necessarily their criteria, are similar, reasonable levels of reliability will be achieved. Although the method is obviously time consuming and logistically cumbersome, it is not as bad as it seems. Once a marker gets into the rhythm, 30 seconds per average high school essay (assuming 45 minutes of writing time) is sufficient. With three scorers, that is less than two minutes per student. The average high school teacher probably spends five minutes on such an essay under normal marking conditions--and produces a much less reliable score.

The problems with holistic scoring are the following:

1. It is only of use in summative evaluation, primarily for classifying and certifying students--it is of no use for individual, instructional purposes;
2. It requires co-operation among teachers, and co-operation takes time and effort;

3. Although holistic scoring is undeniably reliable, questions can be raised about its validity; the following factors may contribute to high impression scores--handwriting, length, and smooth, seamless superficiality; in contrast poorly structured arguments and faulty perceptions may be missed.

In spite of the problems, those who raise them offer no easy solutions. There is great value in including holistic scoring of one or two pieces of writing in the summative evaluation of English two or three times a year. The results should form an important component of the term mark in English.

Consideration should be given to three possible methods of holistic scoring;

Method One: 1 - 10 scoring without standards.
Under this approach, scorers are told merely that they must use all points of the scale. The assumption is that the result will look something like a normal curve, with more 4s, 5s, and 6s than 1s, 2s, 9s, and 10s. When all the scores are collated (and not before), the scorers should collectively determine the cutting points. For example, they might determine that 3 denotes basic competence (i.e. perhaps 50%), 7 denotes honors (say 80%), and 9 denotes excellence (say 90%). Then the other points on the scale will be matched appropriately. The advantage of this approach is that one is likely to get a good spread of marks so that clear judgements can be made. A disadvantage is that, if scorers do not have similar standards, essays marked by two easy markers out of three will do better than comparable essays marked by two hard markers. The problem is compounded by some teachers stressing syntax, others creativity.

Method Two: 1 - 10 scale with standards and criteria.
Under this approach, the scorers agree *in advance* on cut-off points. They also agree on criteria at each level. An agreement might be that 3 denotes minimal literacy--

the essay is understandable but is neither clear nor grammatically accurate. The 5 may denote basic competence--it generally meets the course objectives. A 7 might denote good writing--varied and interesting. A 9 might denote excellence--where the technical, intellectual, and creative aspects of the writing are all considerable. The problem with this approach is that the holistic scheme may break down--markers begin to stop using impression scores and to take too much time. There may also be a tendency for scorers to cluster, with no writing receiving a 1, 2, 9, or 10.

Holistic scoring should not be confused with traditional marking whereby the teacher corrects the errors and then arrives at a score at the end of the essay.

Method Three: 1 - 10 scale with forced scorer equivalence.

This method is designed to compensate for the school's inability to calibrate all the scorers. (Calibration would compensate for the hard and soft markers whose differences will create problems in both Methods One and Two.) Let us suppose that 157 students in grade nine have all written an essay in their June English examination. There are four teachers, one teaching two classes. The teacher with two classes scores all 157 essays, the remaining teachers 105 or 104 giving a total of 471, so that every essay is scored three times. So far, this method is the same as the previous two. However, Table VIII shows how the teachers may be allocated a given mumber of scores in advance. Scorer 1, for example, gives eight essays a score of 10 and eight a score of one. This ensures there is no difference in standard among the different scorers.

Table VIII

Forced Distribution of Scores among Three Scorers

Score	% of Scores	Scorer 1	Scorer 2	Scorer 3	Scorer 4
10	5	8	5	5	5
9	6	9	6	6	6
8	10	16	11	11	10
7	13	20	14	14	14
6	16	25	16	16	17
5	16	26	17	17	17
4	13	20	14	14	14
3	10	16	11	11	10
2	6	9	6	6	6
1	5	8	5	5	5
TOTAL	100%	157	105	105	104

The results will approximate a normal distribution of scores between a minimum of three and a maximum of 30. The percentage of scores column in Table VIII may be used irrespective of the number of essays scored. This scaling process in no way diminishes the responsibility of the four teachers in making a collective judgement about the standard of the essays they have marked. If they decide, for example, that even the highest scoring essay is less than excellent they could give it a mark of say, 75% or B. On no account should these scores be seen as being directly convertible to percentages. No student obtained 29 "out of 30" or 3 "out of 30." In Table IX I have shown what might be a normal conversion of marks with 8 classed as the passing mark and 22 as the honors level. The other equivalents have been arbitrarily interpolated around those fixed points. I have tended to use larger percentage gaps at the extremes assuming the teachers tend to find greater variations in standards at the extremes than in the centre. But that is for the teachers to decide themselves.

Table IX

Sample Conversion of Holistic Scores to Equivalent "Percentages" and Letter Grades

Holistic Score (3 Scorers)	# of Students	%	Grade	Holistic Score (Cont'd.)	# of Students	%	Grade
30	0			16	10	68	C
29	0			15	18	66	C
28	2	95	A	14	17	64	C
27	3	92	A	13	14	62	C
26	2	89	A	12	10	60	C
25	1	86	A	11	6	58	D
24	4	84	A	10	5	56	D
23	2	82	A	9	7	53	D
22	3	80	A	8	3	50	F
21	4	78	B	7	7	47	F
20	4	76	B	6	4	44	F
19	7	74	B	5	0	—	—
18	9	72	B	4	1	40	F
17	12	70	B	3	2	35	F

If one includes time for administrative and committee work, the entire job should still take teachers no more time than they spend on traditional methods of scoring essays. The result will be much more reliable and consequently more acceptable to students and parents.

Using Criteria for Marking Essays

The principle behind holistic scoring is that most teachers have an idea, a feel, for a good essay, based on intuition and implicit criteria related to language, expression, and creativity. Traditionally, teachers have tended to mark essays either with implicit or with explicit criteria. However, whereas holistic scoring produces nothing beyond an impression score, traditional marking produces a variable amount of red ink on the student's essay, presumably related to some criteria.

Generally speaking, holistic scoring is best in a summative situation, when the principal objective is to sort out and rank students in a reliable way. However, criteria are necessary if the marking is to be of assistance for the improvement of performance. It is doubtful how valuable a mass of Catch 22 comments is likely to be when written on an essay. For example, only the superior student who already knows the criteria (e.g. dangling participle, orphan which etc.) will be able to do anything about them. It is more desirable in formative evaluation to mark to a specific set of criteria, and even better if the specific criteria are related to topics recently and actually taught.

A typical set of explicit criteria is:

Ideas	10
Organization	10
Vocabulary, Expression	10
Style	10
Usage	10
Total	50

The use of explicit sets of criteria has more instructional value for the student than have holistic systems. For summative evaluation however, the results are likely to be less reliable than those from impression scoring. Indeed if one wants to provide instructional assistance together with a reliable score, it may be best for the teacher first to read through the essay quickly and give an impression score; after that, the teacher may go over the essay again very carefully marking corrections and making suggestions. Taking those into account in giving an overall score is likely to inhibit the reliability of the scoring. After all, a mass of red ink looks bad. Yet an ambitious writer may stimulate comment from the teacher and make complex mistakes. A much less ambitious and less competent writer may well perform well within his reach.

Teachers differ on whether all mistakes in written English should be corrected. One argument is that marking mistakes is discouraging to the student. Students are thereby given a disincentive to write, particularly if they are dissuaded from using large vocabularies and imaginative expressions. The opposing argument is that failure to mark mistakes reinforces errors. This is particularly harmful if examples of bad English are displayed as good work in the elementary classroom for all to observe. Compromise is probably the best policy. One can mark all the mistakes in a good piece of writing but only the more obvious ones in poor pieces of writing. In this way, marking can be tied into direct remediation. Obviously one has to be careful not to equate, however unintentionally, five marked errors in an excellent essay with five marked errors in a much poorer one. The number of errors is never a good means of determining the value of an essay.

Improving Reliability

Whether one uses holistic scoring or criterion scoring, one must still be interested in improving the reliability, and if possible, the validity of one's marking of essays. Here are some suggestions to achieve improved reliability and validity:

1. Make sure the purpose of the writing is clear;
2. Make sure the type of writing required is clear--first or third person narrative; exposition; development of a thesis, personal or business letter;
3. Restrict the choice of the topic, preferably to no choice at all;
4. Make sure the audience to whom the writing is to be addressed is clear;
5. Give an indication of the desired tone and style;
6. Provide an idea of the form of the writing desired--letter, essay, verse, dialogue, drama;
7. Give some indication of the approximate length required.

By the senior grades, some of those aspects of writing should be readily apparent to the student. A particular purpose will sometimes indicate a particular style and form. However, students should not be penalized on account of the ambiguity of the instructions. The more specific the task the easier it is to measure student's comparative achievement of that task.

In most cases, specificity does not inhibit creativity. Indeed, specificity may well prove to be a spur to creativity. "Write something creative" is unlikely to be a particularly productive exhortation and marking the results in a fair, reliable way will be virtually impossible. On the other hand, the following precise set of instructions may well produce some exciting results from a senior class:

> "Imagine you are Mrs. Rose Hewitt. You have decided to write a farewell letter to your youngest son, Stan, whose eighteenth birthday is today. You are leaving home for an indefinite period. Your husband is an alcoholic and you and he have had a rough life together for the last few years. You love him in spite of the problems he has caused you but you feel you cannot go on any longer. You are not going to tell Stan or his father where you are going or whether you are coming back. You have half an hour to write before you leave to catch the bus. You have a twelfth grade education and you have held only occasional part time jobs."

Note that the above instructions meet all seven requirements but in no way inhibit creativity. To be sure, there will be one student in the class who feels more like "writing a story" or "dreaming up a poem." Let her do that as well as the common assignment. A diversified writing program will give plenty of room for creativity, and students should be encouraged to write at times other than those when a formal assignment is to be marked. Individual, personal writing can often be used as a basis for formative evaluation and development. In marking the Hewitt assignment, the teacher will have to be fairly tolerant of mistakes in language, spelling, and grammar, anticipating the argument that the mistakes are deliberate given that Mrs. Hewitt does not have a high level of education and is under some strain. If one wishes to stress language and spelling, another assignment can be given in which the students are required to write a letter to a rich uncle who operates a fish canning and freezing plant in Prince Rupert, B.C. The students are to request that they be employed in the plant during the summer period.

Assessing Achievement in Subjects other than English

If assessing essays in the English class is difficult, then how much more is it so in geography and history. The history teacher is principally interested in seeing whether students can disentangle the complicated weft and warp of history and proceed to compose a comprehensible fabric. The geography teacher wants learners to assess the balance of the variables which influence geographical phenomena. In addition, both teachers want students to adopt the appropriate, normally expository, mode of writing, in lucid, interesting, and grammatically correct language.

Often, particularly in the lower and middle grades, paragraphs must be required to make pupils exercise the ability to write simple sentences in clear English. Such answers are likely to be mainly factual. Although simple one-word answers usually provide more valid and reliable summative assessments, teachers sensibly include questions demanding written answers because language is "everybody's business."

Usually, essays and paragraphs in the disciplines are marked according to specific criteria. As it is by no means certain that the use of criteria improves reliability (and the reverse seems far more probable) it is all the more important to follow the suggestions in the previous section on improving reliability and validity. Instructions must be specific. One of the commonest errors is to tell students, or lead them to believe, that their own original ideas are what really count and then to proceed to deduct marks for failing to mention certain salient facts. For example, it would be difficult to mark fairly a set of twelfth grade essays written in examination conditions on the topic, "Describe current problems of trade between Canada and the U.S.A." How much needs to be said about the facts? Are they assumed? What if ignorance is inadvertently displayed? What if ignorance is implied but carefully skirted? What if the perceived problems are based on false premises? Does "describe" only mean "relate" or does it mean a request for analysis and commentary?

The following instructions are preferable:

Write a critical essay, about 800 words, on current problems in the trade relations between the U.S.A. and Canada, making sure you include answers to the following questions:

a) What are the major articles of trade between Canada and the U.S.A. today? 5

b) What are the major tariff and non-tariff barriers? 5

c) What reasons are there for people in both countries to object to free trade? 10

d) Describe the effects you would anticipate from the creation of a North American (Canada - U.S.A.) free trade zone. 10

e) Explain why you believe the effects would be those you have described rather than other quite different ones. Evaluate the effects you anticipate from a Canadian point of view. 10

Weight 40; time 60 min.

If marks are to be set aside specifically for organization and expression, students should be told.

That question is likely to produce a more valid evaluation of student's achievement of objectives than is a general essay. A good writer can evade issues, work around ignorance and give poor arguments a surface gloss. Some might argue that such success is clearly the sign of a good writer. That is how advertisers manipulate the public. Even if one accepts such a cynical view of communication, it is surely indefensible to argue that a social scientist should be trained to convince by means of chicanery. That may be what some social scientists do, but it is hardly what we should be training them to do. If that point is accepted, then a skeletal framework will make it much more obvious if a student is unable to cope with some of the important issues.

The reader may well wonder why I advocate the use of criteria in areas such as science and social science but am obviously less enthusiastic in English. There are two reasons for believing a distinction may be desirable. Firstly, an English essay is generally appreciated as a whole and is appropriately viewed in terms of its overall purposes. In contrast, essays in history and science, when written in school, are expected to contain some specific points or analyses. The purposes of rhetorical and scientific essays are certainly not diametrically opposed and may sometimes overlap, but [illegible] believe there are important differences. Secondly, it is a little easier to define specific criteria, and therefore the relative achievement of them, in history than it is in English. The good historian simply does not delude the reader into believing misleading statements. That said, it is still not unreasonable to experiment occasionally with holistic scoring of essays in various subjects other than English. And it is certainly often necessary to apply criteria to English essays, particularly in the formative context. My guess would be that holistic scoring of history essays would tend to reflect English expression more than history--but that is not a question to which any research speaks, as far as I am aware.

Bias in Marking Written English

A teacher's essay marking may be reliable in the sense that marks given one essay are consistent with other essays written by the same individual. Further, they may predict quite well marks given by subsequent teachers. Such evidence does not mean that a given essay marked blind by several teachers will receive comparable marks. Research suggests the contrary (except under the controlled impression scoring system described earlier). One possible explanation of the paradox is that teachers bring a bias to the marking of essays. They already know the "good" students, who are conscientious, persevering, and alert in class. In one sense, this bias is unfair. "Good" students get better marks on essays than they really "deserve." It is possible that bias helps maintain some overall consistency.

I am not suggesting that teachers are unreasonably prejudiced. After all, the high school teacher's marks predict the university teacher's marks quite well--and the latter cannot be aware of the former's biases. Nevertheless, there is much to be said for shared or co-operative marking with every teacher recording her mark without first seeing those given by others. Even if my suggestion of possible bias is totally without foundation, the shared marking scheme will convince students that it is not always favoritism that gives that attractive blond in the front row such consistently high marks.

Evaluation in the Social, Aesthetic, and Emotional Domains

Educational objectives frequently give much prominence to the social, aesthetic, and moral domains. There are many reasons, both legitimate and illegitimate, for this state of affairs, reasons on which I do not intend to dwell in this book.

Although non-intellectual objectives are prominent in vague statements of educational goals, most educational practice in the Canadian classroom appears to be devoted, directly or indirectly, to the achievement of intellectual and cultural objectives. Yet, there are nonetheless, other important objectives and there is usually little point in working towards objectives whose achievement is not to be evaluated. Unfortunately, activities in the social and emotional domains can become a kind of excuse for failure to achieve reasonable standards in the basic skills. "Johnny cannot read and write very well, it is true. But those objectives are rather passé, aren't they? In Empty Minds Elementary School, we like to deal with the whole child. We are more concerned with Johnny's self-actualization and interpersonal relationships than with his spelling." Needless to say there is no formal evaluation in Empty Minds of either "self-actualization" or "interpersonal relationships." There is something of an analogy between the problem of evaluating speaking and listening skills and the problem of evaluating objectives in the social domain. If the school has comparatively little influence on a child's social characteristics (compared with the influence of the home and the peer group), just how much effort should be placed on evaluating a non-event?

The reason most often given for not evaluating social and emotional objectives is that evaluation is impossible. It is certainly not unreasonable to assert that evaluation in these domains is usually less valid and reliable than it is in the intellectual domains. But, it is also true that research studies have generally found small or negligible effects of school programs in these areas, so surely that is all the more reason why schools should avoid extravagant claims.

Let me turn now to some more specific problems and examples in these difficult areas.

Evaluating Attitudes

If we are prepared to assume that it is legitimate for the school, using appropriate methods, to teach some attitudes, then evaluation is not an impossible problem. I am defining an attitude as a semi-permanent, emotional feeling towards an object. It can be placed along a continuum from very favorable to very unfavorable.

The following are favorable attitudes that are frequently considered reasonable ones for schools to develop in their pupils:

1. towards education, school and work;
2. towards useful leisure activities, notably reading;
3. towards other people, particularly people of different backgrounds;
4. towards the natural environment;
5. towards other countries;
6. towards the arts;
7. towards their country.

There is reasonable agreement on the objectives, much less on the means of achieving them. For example, some people who have worked in "multiculturalism" believe that teaching young people about the differences among cultures reinforces prejudices, just as the provision of factual evidence about hallucinogenic drugs may well induce increased usage. It is a cliché that the study of literature in school has inoculated thousands of Canadians against Shakespeare.

Attitudes are not necessarily easily taught. We should be careful then about measuring "something" and then assuming that "something" is produced by the school. Against that, it is reasonable to accept that schools do have some effects on attitudes, and, even if the effect is not great, it is useful for parents and students to know what the attitudes actually are. In a similar way, a medical doctor may tell one about medical problems of whose origin he has no understanding and for which he has no cure.

The easiest way to evaluate attitudes is to ask people what their attitudes are. As almost all pupils know what most of the "right" attitudes are, there is a high premium on honesty. Obviously that means one cannot count attitude test results as part of an overall achievement score. The knowledge that results will count will encourage dishonesty and invalidate the results.

However, if there is good rapport between teacher and student, if the attitude towards the questionnaire itself is favorable and serious, attitude tests can provide useful information to teachers and pupils. The Intermediate Tests include an attitude questionnaire, with sub-tests on areas such as reading, the arts, and education. The various sub-tests consist of sets of forced choice (yes/no or agree/disagree) questions in the form of a quasi-scale. A true scale would have everyone who answers yes (i.e., favorably) to the last and highest question answering favorably all the preceding lower level questions. Similarly, everyone answering favorably in response to the penultimate question would answer favorably the preceding questions but not necessarily the last question.

Honesty, or lack of it, is the greatest problem of the attitude scale. Unfortunately, for evaluation purposes, it is not only problematic that a small minority gives deliberately inaccurate answers. Some people genuinely like to think they have more of the right attitudes than they really have. There is a double problem--attempted deception of the teacher and successful deception of self.

A second way of evaluating students' attitudes is by observing and recording their behavior. There are limits to the poor teacher's time and one can hardly record evidence of every attitude of every pupil, otherwise one would do nothing else. For a few days however, elementary teachers might systematically record evidence of students' attitude to reading. The following five criteria may be used and students rated from 1 (hardly ever demonstrates the desired characteristic) to 5 (demonstrates desired characteristic nearly always) for all of them:

i) reads, when directed at school;
ii) reads, when directed at home;
iii) reads unassigned material;
iv) frequently becomes absorbed in reading;
v) is eager to participate voluntarily in activities demanding reading inside and/or outside the classroom situation.

That scale is taken from Ontario's Assessment Instrument Pool in Reading. Similar scales could be developed for other attitudes.

The two people who most need an assessment of the student's attitudes are the individual pupil and the teacher. Attitudes can and do change. The pupil and the teacher are those most able to change attitudes related to school. Parents must be careful not to adopt a punitive stance towards bad attitudes. Punishing a ten year old for not liking reading is unlikely to provoke much spontaneous joy at the sight of a book. It is better to find books that may actually spark some interest.

Individuals' scores on attitude scales should not be shown to other pupils. It is easy to imagine the sharing of attitudes in a casual way, reinforcing, even propagating, negative effects. Quite apart from that, in the area of feelings one has to be extraordinarily cautious about the invasion of individual privacy.

The teacher need not be overcome by the obvious problems in the evaluation of attitudes. The limitations in the evaluation of essay answers are also considerable--they are just not so obvious. If attitude formation is a genuine objective of the school, it can and should be evaluated.

Evaluating Values

A value, for my purposes, is a deeply held, fairly stable belief which helps to form the individual's conscience and helps provide justification for individual behavior. For this discussion, I am talking principally of such fundamental values as: belief in the family, belief in the golden rule, belief in achievement, belief in truth and honesty, belief in justice, belief in wisdom, belief in the idea of personal property, belief in personal freedom, and belief in self-worth.

Evaluating values is much more complex than evaluating attitudes for at least four reasons: we know even less about whether or how schools teach values; there is some debate even about whether schools should teach values (although there is probably consensus that teaching values is legitimate provided that inappropriate methods of inculcation are not utilized); it is extremely difficult to evaluate an individual's values; and, in some circumstances, it may even be immoral to attempt to evaluate and report on values.

In my view, this is one area, probably the only one, where it is legitimate for teachers to work towards the achievement of objectives without formally evaluating their success. The dangers and difficulties of evaluation seem to outweigh the possible gains. There are two reasons why I do not challenge the legitimacy of teaching values. Firstly, it is quite impossible to have valueless education (totally contradictory to imagine in the case of a system of compulsory schooling). In the second place, I, like most Canadians, agree that school life *should be* predicated on certain basic values.

The school's values, and their inculcation, should be evaluated--but this book is about student evaluation not program evaluation. There are times when the teacher must evaluate a student's values, at least as they are evidenced by behavior. For example, a pupil may consistently lie or cheat. Punishment of the behavior and reporting it to parents constitute a kind of evaluation. Children who are kind, honest, and generous should be praised--once again that praise constitutes a kind of evaluation. Spontaneous, genuine praise is probably the best means of recognizing, evaluating, and rewarding appropriate values. There is only one thing worse than the teacher who never praises students for behavior exemplifying desirable values--and that is the teacher who uses false praise as a weapon and a bribe, thus contributing to the devaluation of the currency of both praise and values. If praise is to be valued, the student must have respect for the person giving it and believe that it is genuine and deserved.

Evaluating Social Behavior

It is hard to draw a line between values and social behavior, which is, in part, a manifestation of those values. (Please note that I am not using a behavioral definition of values--I am not saying that as we behave so we believe. I assume there is sometimes a gap between what we do and what we really think we ought to do. This is very evident in young people who frequently conform guiltily to expectations of the group.) What I have said about values applies to the aspects of social behavior that are related to the basic values of the human being. It is an invasion of privacy for teachers in a pluralistic society to evaluate the value set and associated social behavior of a child. On the other hand, there are, as I have indicated, specific values and specific behaviors, accepted by nearly all citizens, that the teacher must reward. It is obviously unacceptable to criticize a Jew or a Sikh for wearing head covering for religious reasons. It may be perfectly reasonable to disallow a Jamaican immigrant to wear a hat as an act of social defiance.

The teacher must be extremely sensitive in evaluating how well a child gets on with others. For example, introversion is not necessarily less desirable than extroversion. Some children like to work by themselves and there seems to be no reason why they should be classed as social failures on that account. On the other hand, there is an obvious problem if a pupil is isolated and disliked by the group. Even in that case, an unfavorable evaluation of the isolate may well be unfair and unproductive. The problem is as likely to be with the group as with the individual. Once one child criticizes another for being different, others are apt to follow. There are times when introverts simply must co-operate with others. There are other times when they can be given rewarding opportunities for co-operation. The teacher must draw a careful line between the provision of such opportunities and criticizing a child for sometimes wanting to be alone.

Self-respect is another emotional attribute that stems in part from one's interaction with others. This is an attribute, like some of the other basic values, that should not be evaluated in some universal and public manner. To report to pupils and their parents that they are "low in self respect" is not self evidently useful, even assuming the teacher has some harmless, reliable, and valid means of carrying out the evaluation. I must admit to a rather simplistic view of this much talked about problem. The teacher's first job is to see that the individual learns. Learning is helped by appropriate incentives, including the enjoyment of learning, and learning implies success. If young people are happily learning, they are likely to have self-respect within the context of the school. I would not counsel the indirect evaluation of self-respect, although obvious lack of self-respect suggests action on the part of the teacher to provide some enhancement. If self-respect is seen either as being a basic value, to be considered along with the other basic values, or as an aspect of social behavior, extreme caution must be exercised in its evaluation.

There are two areas of social behavior where universal, individual evaluation does appear to be warranted--effort and manners. Teachers usually have beliefs, not necessarily accurate, about how hard their pupils are trying. It is useful for students' efforts to be systematically evaluated and recorded for several reasons. There is probably some general validity in the beliefs of the teachers--after all, some children do very little work and some do a great deal. And effort is one component of success. Pupils and their parents should know if a teacher thinks a student is trying very hard or not trying at all. Similarly, behavior, in the popular sense of the distinction between being courteous and "acting out," should be recorded. Effort and manners are two social attributes that tend to be quite changeable in young people. It is not unknown for a student to try very hard and to behave well in one class, and to be incorrigibly lazy and rude in another. The perception of the teacher is important. Even if the perception is inaccurate, it is important for pupils and parents to know what it is. Subsequent discussion may clear up some misunderstandings. Although it is quite useful to record regularly an overall assessment of pupils' effort and behavior, it is even more valuable to observe and record some of the actual phenomena. Some teachers, for example, provide sessions after school for those pupils who want extra help--and keep attendance at those special sessions. When Jim swears at home he is working terribly hard but Miss Jones refuses to help, it is useful if Miss Jones has an attendance sheet for the voluntary sessions. Similarly, if Jim frequently acts out in class, it is valuable to have a written record of the major events, perhaps even better if they are recorded by Jim himself.

Some Problem Areas for Evaluation

Music and Art

Evaluation in the arts is more than usually confused. One sometimes hears the argument that the arts are personal, experimental subjects and therefore inappropriate for evaluation at all. If the experience of art really were all, then it would be as inappropriate to evaluate pupils' paintings and musical performances as it would be to evaluate the number of flowers they saw on a school outing to a botanical garden.

However, the objectives in the arts are more complex than merely "experience." The point is not redundant however. Evaluation in the arts, as in other subjects, does depend on one's objectives.

Let us assume that objectives in music can be subsumed under the following four categories:

1. Enjoyment of music, i.e. favorable attitude to music;
2. Appreciation of music, i.e. knowledge and understanding and perhaps evaluation of music;
3. Expressive enjoyment of music; i.e. performance;
4. Skills in musical performance.

Let us assume that the objectives in art can be subsumed under the following categories:

1. Enjoyment of art, i.e. favorable attitude to art;
2. Appreciation of art, i.e. knowledge and understanding and perhaps evaluation of art;
3. Expressive enjoyment of art, i.e. enjoyment of painting and sculpting;
4. Artistic skills.

It should be apparent that there are some parallels with the objectives in other subjects--the most obvious analogy being English. Objectives in the second and fourth categories for art and music can surely be evaluated in a relatively straightforward manner, notably so in the case of music. It is not at all difficult to test musical history, theory, and performance. Conservatories, professionally committed to music, carry out fairly sophisticated examinations throughout the country. A superficial examination of these tests makes me think that they are unlikely to be any less reliable than those carried out by most schools in other subject areas. In art too, it is possible, although more difficult, to test knowledge, technique, and artistic production.

Art teachers are often reluctant to declare one water color to be "better" than another as so much depends on personal taste. Yet, the same people will usually not hesitate to categorize the quality of art departments in different schools--and they base their judgement on the products. The situation is somewhat analogous to written English. Marks are not completely reliable. One suggestion is that art teachers try the impression scoring which is so successful in written English. Three different people should give scores on a 1 to 5 scale. The results should then be summed and converted to A to E grades, or even, if absolutely necessary, to "percentages," i.e. places on a 0 to 100 number line.

It is probable that reliability could be improved still further if the three impression scorers were allocated a given number of 1s, 2s, 3s, 4s, and 5s in advance. One division would be into quintiles, i.e. one fifth of the students would get 1 and one fifth 2, etc. Another division, like that suggested earlier for written English, would see 40% of the pupils getting a 3 from each scorer, 20% a 2 and 20% a 4, and 10% each of 1 and 5. There is no right or wrong way. All the impression scoring does is rank students from best to worse. It is still up to the teacher to determine whether the best student deserves an A or a B, the worst a D or an E. If there really is a broad range among the students, as there may be in sixth grade, the use of quintiles will reflect that. If the range is quite narrow, and distinctions much more subjective, as it may be in grade twelve, then the normal curve approach, with more 3s than 1s or 5s may be more appropriate. The same approach could also be taken to assessing performance in debates, in choral and instrumental music, and in drama. One final note on this topic--it should be remembered that three 1s by different scorers do not necessarily mean that the work is unsatisfactory--it all depends on the standards of the class as a whole. Similarly, three 5s do not necessarily denote excellence.

As for attitudes, there is no special difficulty in assessing them with respect to art and music. People tend to be generally honest in these matters provided that no incentive, such as marks that count, is provided for dishonesty. Similarly, it is perfectly possible to evaluate students in terms of enjoyment or artistic expression by observation. Willingness to practise voluntarily is one obvious indicator of enjoyment. Again, I would caution against giving achievement marks for voluntary practice. I once found one of my daughters taking out a musical instrument overnight and leaving it in her locker at school. Why? She found it was an easy way to increase her music mark by way of attitude!

It is sometimes argued that by only counting the final product in an achievement mark, the teacher inhibits enjoyment and creativity. Therefore, it is argued, it is better not to mark formally at all. That is a matter of judgement and opinion rather than one of fact. My judgement is that greater enjoyment comes from doing something successfully than from doing it badly. An emphasis on quality, far from inhibiting creativity and enjoyment, contributes to their development.

Speaking and Listening

The special case of written expression has already been discussed. However, objectives in English usually include four skills--writing, reading, speaking, and listening. I shall not dwell on reading here as the sections on multiple choice and essay questions speak to its evaluation reasonably well.

It may still be useful to make a few more specific comments on speaking and listening, areas that create major problems for the teacher trying to evaluate consistently. It may be as well to clear the decks by stating frankly that speaking and listening do not, and probably will not in the foreseeable future, form a major component of the typical Canadian program in English, the rhetoric of some experts notwithstanding.

There are good reasons for this state of affairs. It is reasonable to assume that young people depend largely on the school for their skill in writing, as the average young person does little writing that is not related to school. To a lesser extent, young people depend on the school for skills in reading, particularly up to the age of ten. In the case of listening, most is carried on outside school hours. As for speaking, only a minute proportion takes place within the school. Even if the 300 minutes of daily instruction were entirely allocated to students' speaking, they could each only have ten minutes of the teacher's time.

In fact, of course, pupils do not spend much of the school day speaking. The creation of small groups is not an adequate answer as there is little possibility of useful, directed instruction without a teacher. Unsupervised speaking may reinforce bad habits as much as good. Once again, there is not too much point in evaluating a non-event. Schools probably have comparatively little effect on most students' speaking and listening. Further, evaluation of speaking is in itself a very time consuming business, which makes heavy reliance on it inefficient.

Nevertheless, within those severe limitations, if objectives include improving listening and speaking skills, and if instructional time is devoted to those tasks, then the achievement should be evaluated. Just as writing should be evaluated according to its function and audience, so should speaking. A formal persuasive address in a debate does not require the same style as a toast to the bride and groom. A distinction must be drawn between the use of colloquial English, often appropriate, and the use of bad English, rarely appropriate. To be sure, the line between the two is fluid and changing, but there remains a distinction. The grunting that sometimes passes for teenage communication should not be reinforced by favorable evaluation in the classroom.

Table X provides a possible analytic marking scheme for argument in an informal debate. Given the minor effect instruction in school is likely to have on informal speech, emphasis is best placed on developing appropriate speech for the classroom, for business, and for formal occasions. Such analytic schemes are particularly useful in formative evaluation.

Table X

Scoring Scheme for an Informal Debate

Criteria	Score
Opening remarks attract attention	5
Argument is developed logically	10
Various points are discrete but connected	10
Voice is assured, clear	10
Mannerisms support the speech rather than distract	5
Content is sound, complete	10
Conclusion summarizes and convinces	10
Rebuttal is logical and picks up major points	10
Rebuttal is convincing	10

Listening is one of the least evaluated skills, yet it is not particularly difficult to carry out. Evaluation of listening skills is also fairly efficient in terms of use of instructional time. In terms of listening comprehension, a passage can be read aloud and students can then be given multiple choice questions, as for a reading passage. Obviously, such an exercise is much more difficult than if the passage is read by and remains in front of the students. The students are paced by the teacher in oral reading; and they cannot go back to check on what they fail to remember. It should be possible to build in questions on the tone of voice and manner of the speaker--providing a good reader brings such qualities out clearly. Recorded tapes are more useful than material read aloud by the speaker as they permit more objective assessment. A group of teachers may together determine the tone and implications of the voice on the tape. A person reading aloud may be quite clear about the intended tone, but the actual communication is sometimes less obvious.

Physical Education

Next to music and art, physical education is the most problematic subject in the curriculum when it comes to evaluation. There seem to be two basic problems:

i) What are the priorities and objectives in physical education?
ii) To what extent should we evaluate the students' progress and to what extent the actual level of achievement?

These are both problems in all subjects but they seem to be particularly acute in this case.

If there is no solution to the two problems, teachers' evaluation will receive little acceptance from students. Dissatisfaction with evaluation, including allegations of unfairness, is a major complaint of students, particularly in secondary schools. It is a fairly easy step from seeing the evaluation as being frivolous and arbitrary to assuming the subject itself has little value. Physical education teachers are often their own worst enemy.

Most members of the public probably expect physical education programs to do two basic things--promote physical fitness and involve students in a variety of physical activities that will carry over into adult leisure life. Physical educators, in contrast, are likely to place a greater emphasis on the development of specific psycho-motor skills, sometimes assuming that skill development in sports will occur of itself and simultaneously promote physical fitness. The difficulty is that there are considerable differences among these three goals and their associated activities. It would seem reasonable to determine in advance the relative weight to be given to the three sets of goals. Students will then know in advance how their marks are to be derived. There are various physical fitness tests available, including the Canadian Test of Physical Fitness and OISE's Intermediate Test. It is not too difficult to devise one's own basic test of cardiovascular endurance, based on a 1000 metre run, and of agility, through shuttle sprints. Similarly teachers can test skill development fairly easily by observation and record results on a five or ten point scale. It is important that class observation take place in a relatively formal context, as distinct from being recorded from memory after the principal has demanded that the term mark sheet be completed. Students, rightly or wrongly, frequently believe that teachers produce marks from the seat of their pants--and accusations of favoritism are particularly rife when it comes to physical education.

Participation in leisure activities is a more difficult problem for the evaluator. A general problem addressed in the next chapter, the legitimacy of adding apples and oranges to arrive at a total mark is germane here. At this point, I shall merely suggest an ad hoc solution; i) if physical fitness is a major goal (and surely it should be), give it a major proportion of the weight - 50% to 60%; ii) give perhaps 20% of the weight to specific psychomotor skills; iii) give the remainder to "performance in sports activities."

That last will rely heavily on the actual level of achievement, but if sports are selected that require relatively few highly specialized skills, e.g. soccer, then effort and attitude will obviously contribute greatly but indirectly to success.

That leads us directly to the second problem--to what extent should absolute achievement be marked compared with progress? That this is a matter of judgement does not make all approaches equally defensible. Natural ability and previous background in the subject are enormously important to success in virtually all areas. High intelligence coupled with a supportive home background make success in mathematics extremely probable. One does not give an A to the student who could neither add nor subtract when he arrived in grade nine just because he has learned both skills by the end of grade twelve (unless the student is clearly identified as being a member of a special class with such clearly stated and limited objectives). It is therefore inconsistent with the rest of high school evaluation practices to mark students in physical education mainly on the progress they apparently make in the course. I say "apparently" because it is neither unknown nor difficult for students to simulate incompetence at the beginning of a course if they know they are to be marked on progress.

It will be argued that if a student enters a course already brimming with fitness and physical competence, that student can coast and still get the top mark. The same is theoretically true in any subject. The idealistic answer to that well-founded complaint is that secondary schools should be organized on a basis of competence rather than on a basis of time. With that system, one simply would not have students taking a course whose objectives they had already achieved. The practical answer is twofold. Firstly, the alternative cannot be made to work fairly or reliably. How can one compare the progress of the expert badminton player with the progress of the beginner in any reasonable way? Secondly, the teacher should adapt the program so that it stretches even the most accomplished member of the class. In that way our prize physical specimen will have to work for that mark of 100%, even though 90% may be fairly easily achieved. That brings us back to the *nice to know* idea. The *nice to know* category might account for the final 15% of the mark, thus challenging even the most capable student.

In short, physical education teachers should come to terms with their objectives and mark accordingly. Marks, as in other subjects, should reflect relative, final achievement, but that does not mean that the talented should be allowed to coast.

Mathematics

Mathematics and the sciences are usually not seen as problematic areas for evaluation. Generally, students have few complaints in these areas. Answers are seen to be either right or wrong.

However, it would be a mistake to think marks in mathematics are necessarily reliable. Two good teachers of mathematics could easily give the same student 50% and 90%, depending, for example, on the proportion of marks given for presentation and the proportion given for the right answer.

I am not recommending that math teachers rely entirely on multiple choice questions with the concomitant emphasis on the right answer. It is perfectly reasonable to expect students to understand the method and approach. Emphasis should be placed on method, particularly in formative situations, and greater emphasis placed on correct answers in summative. Usually, the lesser the emphasis on method and the greater the emphasis on correct answers, the more reliable will be the marking. A good idea for a summative examination might be a specific subsection dealing directly with presentation, method, and format, with the major section being written in a multiple choice format assessing ability to obtain right answers.

Projects

Projects often form a significant part of teachers' marks, particularly in elementary schools. Unfortunately, the evaluation of projects can easily reinforce the achievement of objectives very different from those used to justify the project in the first place. Often, for example, projects are given a single total mark that principally reflects presentation, neatness, attractiveness, and effort. Worse still, the mark may reflect external assistance or the ability to copy.

If, in geography, the teacher merely asks students to pick a South American country and do a project on it in teams of three, the result is likely to be an uninspiring collection of copied maps, cuttings from National Geographics (to the distress of the school librarian) and regurgitated facts from encyclopedias. It is extremely doubtful whether any worthwhile objective is achieved through such an exercise. The evaluation of such projects, based on the volume and neatness of the product, fails to meet the most basic criterion of evaluation--that it reflects some worthwhile agreed upon objective.

Valuable projects require a good deal of work on the part of both teachers and pupils. Evaluation should, as always, be based on justifiable objectives. A project in science can be justified if students obtain some experience of working scientifically. Let us, for the sake of illustration, assume that science consists of the following kinds of activities: observation and description; classification; development of hypotheses; experimentation; induction; deduction; logical analysis; creative synthesis; generalization--all of these being within the disciplines constituting the physical and life sciences. It follows that a science project should be evaluated in terms of the quality of work in one or more of those activities. That, in turn, implies detailed specification about what is, and, especially important, what is not wanted from the project. A sixth grade teacher may object, "Oh that is too difficult for ten year olds. OISE professors don't understand what ten year olds are like." If "all that is too difficult," one should avoid science projects (and perhaps science itself). Giving an A in science for a project that bears no relation to the stuff of science, is miseducative. However, I believe most young people, even in sixth grade, can at least observe, describe, and classify plants, fish, minerals. Most can probably figure out simple experiments and record the results. The point being made is that evaluation must reflect the central objectives.

Presumably, the use of projects is justified in terms of learning. In that case, it should be justified in terms of achieving appropriate and relevant objectives.

Summary

This chapter has provided a general overview of the day-to-day problems facing the teacher in the classroom with respect to evaluation. Intellectual skills have been examined in terms of the conventional categories of knowledge, understanding, and thinking. The suggestion is made that teachers plan their teaching and their evaluation according to a three-way system of *must know*, *should know*, and *nice to know*. There will often be more "knowledge" than "thinking" within the "must know" category, but that is not necessarily so at higher grade levels.

Various types of test questions have been described. The particular advantage of multiple choice questions for evaluating sophisticated analytical objectives has been stressed. The advantages of essay writing are in evaluating, synthesis, and creativity, providing that the major hurdles of reliability are overcome. Essay writing is particularly important if writing itself is to be adequately encouraged and improved.

Problems of evaluating objectives related to emotional, social, and aesthetic domains have been described. Attitudes are relatively easy to evaluate, but values and social characteristics are much more difficult. Particular care must be taken to avoid unjustifiable invasion of privacy.

Problems related to difficult areas, such as art, music, physical education, oral English, and projects, can be overcome, providing the objectives are clear. Difficulties in evaluation very often reflect a failure to decide what one is actually supposed to be teaching. Although one should not completely avoid evaluating those things that have been learned outside the school rather than within, there should be some limitation of the energy put into evaluating those things over which the school has almost no influence.

4 PLANNING AND ORGANIZING STUDENT EVALUATION

Chapter III was intended principally for the classroom teacher, who has the daily task of evaluating tens even hundreds of students on almost as many objectives. Many of the difficulties discussed in that chapter raise problems whose solution lies beyond the power of the classroom teacher no matter how decentralized are the school and the school system. Are the objectives clear and reasonable? Do values, attitudes, and intellectual achievements all have to be lumped together in a single percentage for the computer? Is a mark in art provided on the home report? It is with these questions of policy that this chapter is concerned. At first sight, it might seem that this chapter will be of lesser interest to teachers or parents, who may see decisions in these areas as being beyond their control and perhaps even being administrative rather than pedagogical. I do not think that is the case. Increasingly, teachers and parents are being involved and are involving themselves in decisions of this kind. As was the case in earlier chapters, not all readers will agree with the suggestions made in this chapter, but all involved in education should be aware of the issues and the problems.

Allocating Marks in the School and Classroom

Individual teachers usually have a great deal of autonomy in the allocation of marks. After all, in the final analysis it is their professional judgement of a pupil's achievements that is most important. Nevertheless, it should be obvious that the allocation must take place within some kind of framework of policy. Even the most extreme advocate of "teachers' rights" will surely agree that the teacher cannot have the right to be unprofessional--to give high marks to pretty girls and low marks to gauche boys; to favor pupils who bring presents; to give high marks to the children of friends. Obviously then, there has to be some definition of what are professional ways of allocating marks.

Most teachers will probably agree that within a school, perhaps even within a school system, there has to be an agreed, consensual framework. Marks are, for better or for worse, one of the most significant incentive systems in the school--easily the most important at the secondary level. If different teachers in the same subject use entirely different methods, so that Sue's identical twin sister Rachel gets an 'F' for work that earns Sue an 'A,' students and parents soon lose confidence in the integrity of the system and even the integrity of individual teachers.

Let us examine some of the problems and some of the solutions with respect to mark allocation. Let us assume, in all cases, that the teacher is adopting, at least implicitly, the *must know, should know,* and *nice to know* categories of learning and that the objectives of having all students achieve basic competence and the best students achieve excellence are given equivalent priority.

Scaling

Although this system is scarcely, if at all, used by Canadian classroom teachers today, it is worth examining its rationale because it clarifies some of the problems in mark allocation. Scaling is a mathematical process, rather complex and time consuming without a computer, by which marks within a class or group of classes are arranged according to some predetermined pattern, usually based on a normal distribution. A simple variant is called rank scaling, a type used experimentally in a school of which I was principal. With rank scaling, the first student in mathematics (or French or English) might receive 100, the second 97, the fifth 84, the tenth 78, and so on. In our case, we had tables published for classes of different size and for different levels (enriched, regular, and basic). Teachers were able to vary marks at the extremes to reduce or increase the incidence of very high marks and of failures. Otherwise, they ranked the students they were teaching and gave them a mark from the table.

The teachers in the school agreed to adopt this scheme experimentally because the problems of the existing laissez-faire situation were readily apparent to everyone. Some traditional teachers thought education consisted essentially of learning a set number of facts. They believed a mark of 90% indicated that 90% of the facts had been learned. Their marks tended to be rather high. Their tests were entirely factual. Other teachers thought that thinking and understanding were more important. Their marks tended to be lower. The students with only factual knowledge might do quite badly or even fail. Other teachers thought that students should be marked according to the teacher's assessment of their classroom experiences. Their marks tended to be high but arbitrary and difficult to defend. There were several English immigrant teachers who thought a mark of 90% preposterous, particularly in the subject of English, and whose marks tended to be generally lower than those given by native-born Canadians. Scaling meant that the best students got the same marks whichever course they happened to be enrolled in and whichever teacher they happened to be assigned. In this sense, scaling was much fairer than the old system.

There are several problems with scaling, the last one of which is fundamental. Scaling means work, even if one uses rank scaling tables. It removes most of the teacher's right to judge the absolute value of work--leaving only the right to rank the students relatively. It totally obscures any real differences of achievement among different classes. For example, let us consider two classes taking the same course. Teacher A is excellent. The students work hard and achieve a great deal. Teacher B is mediocre. The students play around and achieve very little. Yet with scaling, the results of the two classes are identical. (It must be remembered, however, that in the laissez-faire system teacher B may well give even higher marks than teacher A, in the hope the high marks will conceal the evidence of lack of learning--at least the parents will not complain.) Obviously, if marks are scaled, much of the incentive to teachers is removed--if we assume that teachers are rewarded by high achievement legitimated by high marks on worthwhile tests. Scaling can also very easily mask the significance of the *must know*, *should know*, and *nice to know* categorization--the results are the same whatever happens to be tested. Those problems are considerable enough but there is another one that overrides them all.

Students quickly come to realize that, collectively, the amount of work they do has no effect on the range of marks. Obviously, it is a short step from that realization to an agreement not to be a "curve-bender." Scaling provides the incentive for students to band together to beat the system--and the teacher.

Scaling is probably never going to be used by any school or system on a long-term basis on account of those endemic problems. It is, of course, a basic element of standardized tests. It is also used in provincial and state-wide examinations to ensure general consistency from year to year.

The value of this excursion lies in the illustration of the laissez-faire system. In many Canadian provinces school marks are almost the only basis for movement from high school to college or university. Access to narrowly restricted programs and to scholarships is determined by some average of high school marks. Obviously the range of standards from school to school is immense. Thus the student is "lucky" who attends a school with low standards and high marks and "unlucky" who attends a school with high standards and low marks. The second problem, of the variation of marks within the schools, is a little less obvious and I shall return to this later.

Variations of Marks and Standards among Schools

This problem is most serious at the secondary level. Elementary schools can easily use readily available standardized tests to ensure comparability among students as they pass to senior public school, junior high school, and high school. No such easy solution is available for secondary schools, yet the consequences of not dealing with the problems are much graver. Whereas elementary students usually pass freely to secondary school, there are strict enrolment limits to popular university and college programs, such as pharmacy, nursing, engineering, and accounting.

One way to solve the problem within a school district is to have a district examination count for part of the school mark. The disadvantage of that idea is that, unless the exam counts for 80% or more of the total mark, extreme variations among schools will still outweigh the effect of the examination. Research in Ontario recently showed that, at least in one region of the province, the majority of high schools' standards only varied moderately, but the extremes were very great indeed. It is just those extremes that are not looked after by, say, a 50:50 split between school and central examinations. And for the remainder, the result may not be worth the effort.

More complex, but more desirable, is a solution whereby a single examination is given to all students in a given course in all schools. The students' marks from their teachers are then scaled on the basis of the class's performance on the examination. The actual examination may still count little. In practice the exam should be given a weight sufficient for students to have an incentive to perform well and take it seriously.

Table XI shows a simulated set of marks providing a picture of what happens when marks are scaled. Class A has done comparatively poorly on the examination but has a very high set of school marks. The scaling provides a mark consistent with that of the other classes, the average scale marks being therefore lower than those of the other classes. However, student A2, with the highest school mark, still gets the highest scaled mark in the class, because the scaling reflects the school's ranking. (In effect, the ranking within the class is determined 70% by the school and 30% by the exam, if that is the distribution decided upon.) This process thus lets the examination set the absolute level and range of marks for the class, but the ranking of students within the range reflects the school's judgement more than the exam's. Class B is least affected by the scaling as its school marks and exam marks are both intermediate. Class C presumably was marked in school according to high standards, receiving the lowest school marks and the highest exam marks. Its scaled marks are higher than the class marks, although, generally speaking, the students classified best by the school still get the best results.

In sum, some kind of scaling is desirable when one wants an overall external standard to be imposed. Scaling still permits teachers to determine the relevant ranking of students within the class or school. Scaling is valuable among schools because it means they all compete on an equal footing.

Variation of Marks within Schools

Mark variations within the secondary school come from a number of sources:

a) some classes have better students than others;

b) some classes have better teachers than others;

c) some teachers mark harder than others, perhaps by giving tougher assignments and tests;

d) some subjects traditionally receive different mark distributions from others.

Table XI

Estimate of Marks Scaled on Basis of Central Examination

Class A				Class B				Class C			
Pupil	(100) School Mark	(30) Exam Mark	Scaled(100) School Mark	Pupil	(100) School Mark	(30) Exam Mark	Scaled(100) School Mark	Pupil	(100) School Mark	(30) Exam Mark	Scaled(100) School Mark
A 1	95	16	73	B 1	93	28	95	C 1	49	15	58
A 2	97	12	75	B 2	37	14	56	C 2	67	17	72
A 3	89	17	70	B 3	94	30	98	C 3	78	20	92
A 4	73	9	58	B 4	55	16	68	C 4	72	21	85
A 5	84	12	67	B 5	72	17	80	C 5	86	29	98
A 6	98	14	76	B 6	65	19	72	C 6	58	17	63
A 7	69	3	46	B 7	84	25	88	C 7	74	23	88
A 8	82	10	64	B 8	46	13	65	C 8	73	20	87
	Mean of Whole Class: 88.7 Mean School Mark 13.8 Mean Exam Mark				Mean of Whole Class: 75.3 Mean School Mark 19.3 Mean Exam Mark				Mean of Whole Class: 73.6 Mean School Mark 20.6 Mean Exam Mark		

The first two causes of mark variation (a and b) are more matters of general administration and pedagogy than evaluation. Better students perform better than worse students so geography marks are likely to be higher than history marks if most of the bright students for some reason have chosen the former over the latter. Similarly, some teachers are more effective than others. The good administrator tries to help improve the poor teacher, but raising marks to compensate for inadequate instruction is not a particularly helpful solution. There may be an exception to that generalization. Suppose a class is in its final year before graduation and has received very poor instruction in mathematics. If marks are based only on achievement in that course, even the brighter students may fail to gain access to the university and program of their choice. In extreme circumstances, it may be justifiable for the principal to replace the achievement marks with estimates based on overall tests of mathematical competence, always provided that the students know what is happening and are not led to believe that they have learned what they have not been taught. A sensitive administrator should be able to do this without humiliating the teacher at the same time. Ineffective teachers sometimes raise marks, or lower standards, in compensation for their pedagogy. The administrator and the parent should avoid being misled by apparently acceptable marks.

Let me turn now to the greatest problem: variations from teacher to teacher in the same course (c). The administrator has to maintain a precarious balance between fairness to students and pedagogical autonomy for teachers. The extremes are easy to categorize, a reasonable compromise more difficult. Few would argue that, in order to be fair, every student taking the same course must be given exactly the same assignments marked by exactly the same procedures. Similarly, few would give secondary teachers the individual right to determine all marks totally independently of any guiding policy. And the desired policy must go beyond personal ethics.

Scaling is one answer to the problem, but, generally speaking, the cure is worse than the disease. Much of the problem can be mitigated, even obviated, if program content and methods of evaluation are agreed upon in advance. There should be clear school policies on such matters as absenteeism and failure to complete assignments, problems to which I shall return later. Teachers of English may agree on a program of combined holistic impression scoring of the kind outlined in Chapter III. Teachers of mathematics may agree on the type of format they want for mathematical problem solving. They may agree on the weight to be given to the answer and the weight to be given to the method. Teachers of art can adopt a variant of the written English impression scoring method to evaluate works in different media. Teachers of music can agree on the weight to be given instrumental performance, understanding of theory and aesthetic appreciation.

It is failure to come to such formal agreements that causes a large proportion of the discrepancy among teachers. One teacher bases most of the mark on home assignments, another on class tests. One teacher marks attitudes. Another marks progress. Yet another marks achievement. These problems can and should be solved in advance.

However, let us assume that reasonable steps have been taken to compile common policies and practices. The allocation of marks among tests, examinations, and home assignments has been determined. The distribution of marks for process and product, in mathematics and science, has been determined. Physical education marks for fitness, motor skills, and performance in sports have been laid down. The relative importance of musical expression and technique, and their overall importance compared with theory and aesthetics, are known. There still remains the problem that some teachers have different standards from others. For a given essay, teachers may disagree on the absolute value of the quality of argumentation. On a historical essay, teachers may disagree about the quality of the background research and its relevance to the issue.

Thus, even with the best possible policies--agreed upon, disseminated, implemented--there will still be substantial disagreement, particularly in areas requiring a higher degree of subjective, critical judgement. This problem is of a very different order from that of the systematic discrepancies discussed earlier. It is unfair if one teacher looks for correct answers, another for appropriate methods. It is unfair if one teacher gives all the marks on the basis of physical fitness, while another gives them all on the basis of skill in sport. In that kind of circumstance, students can legitimately complain that their marks are overly dependent on the capriciousness of fate. They understandably become cynical and believe that passage through school is partly a game of chance and partly a contest between them and the teachers.

The same objection could be extended to the situation where some teachers systematically set higher standards than others. For example, I was once principal of a new secondary school whose English department was staffed variously with people of American, British, and Canadian origin. The Americans marked most generously, 95% was a good mark for them, the British least generously, 75% was a good mark for them, and the native-born were the compromisers. The staff soon recognized need for a good Canadian compromise and the marks came into line!

However, the problem is less clear where the differences are ones of idiosyncratic rather than systematic judgement. The differences may still be considerable, but there is no overall difference either in the average mark given or in the range of marks. There can be, and probably should be, no simple solution to this problem. An arbitrary decision about what is good or bad, where reasonable experts in the field differ, is not usually desirable. However, the problem can be alleviated, particularly in large schools, by giving students more opportunity to change classes, by co-operative marking, by common objective tests, and by impression, holistic scoring of essays. In the final analysis, quality and excellence remain individual judgements of experts, and there will always be some difference among them. What is necessary is that there be an agreement on the search for excellence.

Variations from Subject to Subject

Variations from subject to subject are not necessarily bad. If students have to choose between biology and chemistry, the biology class may well have less able students than the chemistry class. It would hardly be sensible to try to make the pattern of marks similar. The principal who does not like to "see failure rates above 20%" will inadvertently encourage lower standards in biology and higher standards in chemistry, a trend that the natural selection of students has already encouraged.

The biology teachers are likely to try frantically to keep the failure rate under 20% while a chemistry teacher is given a licence to weed out the "non-scientist" or make the course even more rigorous. A more reasonable administrative expectation would be for the chemistry marks to be somewhat higher than the biology marks.

A matter for careful consideration is the possibility of there being systematic variation from subject to subject unrelated to student ability. For example, my experience is that the range (not the level) of marks in English and humanities is lower than the range in mathematics and sciences. To check that experience, I obtained average school marks from data collected in Ontario in 1977 for the Interface Project (a Ministry of Education project examining movement from school to college and university). The results are shown in Table XII. Clearly, as predicted, there are important differences among this random sample of secondary schools between mark variation (standard deviation) in English and French on the one hand and variation in physics and math on the other.

Table XII

Mean Standard Deviation in Various Subjects in Ontario Secondary Schools

	Grade 12				Grade 13		
	All Students	Over 49*	N		All Students	Over 49*	N
Advanced English	11.38	9.47	48	English	10.90	9.58	51
General English	10.27	8.88	45	French	10.81	10.29	50
Advanced French	12.11	10.59	47				
Advanced Mathematics	13.87	12.28	46	Physics	13.15	11.50	51
General Mathematics	12.95	10.99	42	Calculus, Relations & Functions	14.30	12.19	50
Technical, Business Mathematics	12.58	10.91	28				

*Counting marks 50% or over only

That statement holds whether one looks at marks of all students or confines one's attention to passing marks over 49%. The separate analysis of marks 50% and over shows that the discrepancies between math and English are not explicable solely by the presence of a larger proportion of extremely low marks in mathematics.

There are reasons why there are differences of this kind. It may be argued that there are coherent differences in the nature of the subjects. Questions in mathematics have right and wrong answers whereas there can be no perfect English essay. Certainly, those differences are not unique to Canada. Teachers everywhere tend to spread marks less in the first language and more in mathematics and science. One could argue that this universal tendency shows that the distribution of competence in mathematics is greater than it is in English--the marks just reflect reality.

That statement opens up a Pandora's box of measurement problems. In what sense can one say that one variable varies more than another? By what criterion can one determine whether height varies more than weight, or that love of ice cream varies more than love of flowers? Comparisons of differences assume identical measuring scales. If one looks at standardized test results in middle grades, one could make a case for precisely the reverse--that students vary in reading more than in mathematics. It is not uncommon for a seventh grader to read at a "tenth grade level," very rare for such a pupil to perform in mathematics at that level. In return, one can argue that growth in mathematics is much more rapid than in reading over those three years--but that too, brings us back to the problem of such comparisons. How does one determine whether a whale grows faster than a primrose?

I have probably exhausted the reader's patience by this time on a question to which there can be no conclusive answer. Teachers are clearly not necessarily foolish because they give less varied marks in English than in mathematics. The poor reader may be wondering why on earth it matters.

The problem stems from the fact that marks are very important, particularly in the later grades of high school. They determine the relative academic competence of students. They determine those eligible for college. They determine a group of failures. They determine students' sense of self-respect in terms of their academic work. They determine students' feelings about their relative competence in different subjects.

The different distributions of marks influence all these matters. It is easier, statistically speaking, both to excel and to fail abysmally in mathematics than it is in English. A bright, all round student, is well advised to concentrate on mathematics and some sciences in order to achieve the highest possible average in the Canadian school. That fact helps encourage the more intelligent students to concentrate on those subjects, less intelligent students to choose English and social sciences, in which it is difficult to fail. It means also that, statistically, poor English is much less of a barrier to successful completion of school than is poor mathematics, if both are compulsory or expected courses. It is rare for high school students to get the message that they are brilliant in English, more common to hear that about science. Those with a scientific bent achieve scholarships more frequently than those with a tendency to the arts. Bright students congregate in universities' science programs, duller ones in the social sciences.

This move to the sciences is supported by other factors--a growing dependence on technology and high salary levels in medicine and engineering. However, I would argue there are problems when our mark structure systematically gives students a biased view of their competence, and, when, for example, a student with a superior but not outstanding average, say 85%, in twelfth grade believes the marks of 89% and 87% in physics and chemistry denote a better performance than marks of 83% and 81% in English and history. Using the Ontario sample from the Interface study a mark of 88% in advanced math is statistically equal to a mark of 82% in advanced English.

Although the mark structure is international, the problems created are peculiarly Canadian. In the U.S.A., competition among high school graduates is partly moderated by standardized entrance tests, on which the mark distribution in different subjects is equal. In Britain, examinations are administered nationally and grades are given in the form of letters, with school marks being irrelevant. Those Canadian provinces that depend exclusively on school marks for passage from high school to university and college are unique among major western countries in their failure to address these problems.

The easiest solution to the problem is in fact use of the letter grade rather than the percentage system. If students' marks on a course are displayed on a continuum from high to low, the teachers, in consultation with the administration, can determine cut offs between A and B, between B and C, and so on. English teachers are, in those circumstances, likely to award just as many As as mathematics teachers. For some reason, they have no aversion to an A, but they do dislike a score of 100. In schools where such a change is not possible, perhaps because of school board or provincial policy, education and persuasion may be the only weapons. It is probably easier to change the teachers whose marks disadvantage their subject than those whose marks provide their subject with an advantage. School marks in different courses should be publicly compared and justified. It should be possible to persuade English and history teachers that they are hurting themselves as well as their students if they pass everybody without specifying clearly some minimum level of competence, and that they are hurting the most able students if they consistently give them marks that mislead them into thinking that they are not very good in the subject.

More General Problems of Evaluation Policy

Levels and Streams

Levels

The typical Canadian secondary school provides courses in some subjects at different levels. Courses may be given at an advanced level to prepare students for entry into university while a second type of course, at a general level, is intended to prepare students who are going directly to employment. The different levels can easily lead to confusion if students, parents, teachers, university registrars, and employers do not understand the difference. A grade of B, or 85%, does not, and should not, correspond to the same level of achievement in both advanced and general English. What is necessary is the spelling out of the differences. In many cases, a competence level equivalent to, say, an A in a general course can be considered equivalent to just passing in an advanced course; and that does not imply that one course is a watered down version of the other.

As one progresses through the secondary school, the levels will probably diverge greatly in content, particularly in subjects other than English, thus making any kind of direct equivalence impossible. However, absurd restrictions should still be avoided. Successful completion of advanced eleventh grade mathematics should normally be a sufficient qualification for entry into general twelfth grade business mathematics, although the converse would not hold. It is unlikely that the eleventh grade business program will have included much that cannot be quickly reviewed by the advanced students.

In small schools, there may very well be more than one level of course in the same class. This necessitates a decision being made as to whether the class is following one program or two. If the class is following one program, then there should be no differentiation of marking among class members. One sometimes hears the argument that it is unfair to mark a slow student and a student who knows little English in competition with gifted students. That argument is pernicious. If anything is unfair (and that depends on what students want, and what the school should reasonably provide) it is the infliction of an unsuitable course on the students. If, rightly or wrongly, students are taking an inappropriate course then it is unjustifiable to give them high marks to compensate. "Adjusting" marks in those circumstances is no charity to the student--it is a protective device for the school by which it can conceal its failure to provide appropriate programs.

If the decision is made that only one course can be offered to the whole class, then all students should be marked on the same standard. However, teachers should be particularly careful to heed the advice given earlier and ensure that no more than minimal competence, and no less, is required for a passing mark. This in no way prevents the teacher from giving differentiated assignments permitting genuine discrimination between the student making 85% and the student making 95%.

Careful consideration should be given to offering two or more programs to a single class in a small school. Secondary teachers, who are generally unaccustomed to such an idea, should remember that elementary classes frequently have two, three, or more groups operating at the same time. If general and advanced English courses are made available to the same class, then obviously different marking schemes should be used for each, and the minimum standard for one should not be the same as the minimum standard for the other. From a pedagogical point of view, it is probably better to emphasize the difference in content of the two courses rather than difference in standards, of which the students will, in any case, be well aware.

Streaming (and the problem of movement from intermediate to secondary school)

Streams, tracks, and levels are terms used without complete clarity. I am using the term *level* to refer to the situation where there are different programs in the same subject at the same grade. *Track* is an American term used to refer to a situation where students are forced to choose all or most of their courses at the same level. Students thereby become *tracked*--as college preparatory or vocational, perhaps. *Stream* is a similar term used particularly in Britain to refer to a situation where classes are graded from most advanced to least advanced. Individual scheduling is rare in Britain and students are typically placed for most subjects in either the A class, the B class, or the C class. *Setting* is sometimes used so that students in the A class for French can be placed in the C class for biology--this is achieved by scheduling several classes of the same subject at the same time. *Setting* differs significantly from the common system of *grouping* in the elementary school. Under grouping, students, usually from a single class, are put together according to their competence in a particular subject, most frequently in reading and mathematics.

I would advocate the application of the following general principle. If students in different classes, or groups, are following the same program they should be marked on the same basis. Thus, if classes 7A and 7B are following the same program, they should be evaluated by the same procedures and standards. To those who argue that the slowest students in 7B should be marked more generously, I would respond that the problem is one of instruction rather than of evaluation. My view is that there is not much point in segregating pupils on the basis of ability or competence if they still follow the same program. Certainly, it is confusing and misleading to evaluate them on different bases.

However, if 7A is an accelerated or above average class, perhaps taking three years mathematics in two, or including a much more demanding range of reading material, then it makes sense to mark the class on the basis of expected achievement in that particular class. In that case, parents and students should both understand that the work is different and that the grade of C in 7A is not equivalent to C in 7B.

The situation is not always so cut and dried. However, students and parents should always know the program being followed, the objectives of the program, and the means by which progress will be evaluated. Sometimes one has a situation where students are partially *streamed*. This frequently happens at the intermediate or junior high school where principals are scared of the rigidity of streaming and yet find the range of competence in a heterogeneous class virtually unmanageable. Students are sometimes (most unfairly) put into different classes depending on whether or not they have selected, for example, an optional course in instrumental music. The instrumental class becomes the de facto advanced class in all the other subjects, with particular hardships for the bright child who does not want or cannot afford instrumental music. Sometimes, students are streamed somewhat informally, with the middle classes being undifferentiated and the top and bottom classes being more clearly *enriched* and *remedial* respectively.

Problems of evaluation in these circumstances are frequently symptoms of poorly thought out programs. If successful students from all classes are to proceed to high school on the same basis, there seems to be little justification for separating them in the first place. If an *enriched* class is to receive a substantially different and more demanding program, then criteria for admission should be clear and public. In that situation, different methods of evaluation are clearly appropriate, providing that the receiving high school understands the difference. Needless to say, if the junior high school principal only has some vague feeling that the so-called *enriched* class is marked "a bit harder," then the secondary school administrator can hardly be faulted for not having differentiated placement and admission policies. If the *enriched* class is really an advanced class, it is reasonable to expect the secondary school to provide exemption from work already covered. Exemption is likely to be based on satisfactory proof of competence.

Some form of common assessment, possibly in the form of standardized tests, becomes essential in situations where intermediate schools have different levels or streams. The more vaguely their levels are determined, the greater the need for and dependence on common, external assessment. The greater the number of feeder schools, the greater the variation in their programs of instruction, the greater the need for external assessment.

In summary, if programs are essentially identical, the methods of evaluation should be the same and so should the standards. If programs are differentiated, then it is reasonable to change the methods of evaluation, and, in many cases, the standards. However, in the situation of the intermediate or junior high school feeding students to high school, one has to be careful defining minimum competence. Suppose the junior high school has three levels of program in ninth grade mathematics feeding into three not necessarily complementary high school levels. The minimum pass standard in the different junior high levels should be either related in some clear fashion or identical. For example, it is anomalous if a student who passes *regular* ninth grade mathematics continues into *advanced* tenth grade mathematics in high school, when, at the same time, a student who fails *enriched* ninth grade mathematics is told to choose between repeating grade nine and taking *general* tenth grade mathematics. That latter student might easily be better than the nominally successful student from the regular class. Either the pass mark should be the same in all streams or minimum marks should be differentiated and related.

Ideally, the high school should set its minimum levels of competence for its various courses in co-operation with the feeder schools. Marks in the junior high can then be related to high school entry standards. For example, a mark of 40% in *enriched* ninth grade mathematics might be acceptable for entry into *advanced* tenth grade mathematics, while 60% might be necessary for a student proceeding from a *regular* ninth grade mathematics program.

In many provinces, students have the right to choose their level of course in high school providing they generally meet the minimum requirements for success in the last year of intermediate or junior high school. That situation, which I consider a desirable one, does not in any way make my suggestions for relating high school and junior high standards irrelevant. On the contrary, it makes that relationship even more desirable. Too many students enter high school programs for which they are unsuited and poorly prepared. They are less likely to make unwise choices if they can see exactly which competencies they lack. Success in junior high school should be related to minimum demands in high school courses.

This is not the place to discuss the organization of high school programs, an extremely complex topic. Even so, I should make it clear that I am not advocating the setting up of a single standard which, in effect, becomes a high school entrance examination. I am suggesting that there should be some continuity in students' programs. If high school programs are ordered sequentially, students can be placed in the appropriate spot in the sequence--and still retain their right to choose between advanced and general levels, if those levels still remain necessary when appropriate sequences have been developed. The right to choose is one thing, the right to accurate information on achievement is another. If marks, streams, grade levels are totally confused, it is not surprising if students and their parents sometimes make unwise choices.

In Canada, large numbers of students embark on college preparatory programs. Yet, fewer, only about one fifth, enter university, and about one third take any form of post-secondary education. Sound evaluation techniques will give students better information about their strengths and weaknesses, and make intelligent choices easier.

Movement through the Grades

The problem of streaming is particularly difficult in North America because it involves the grafting of an exotic species on an indigenous root. Traditionally, North American schools have promoted students from grade to grade on the basis of their average achievement.

That process has been subjected to stress and change at the secondary level with the advent of individual, computerized scheduling and at the elementary level with the idea of *continuous progress*. Nevertheless, the core concept remains in all Canadian provinces today. In the typical secondary school today, students are still promoted from grade to grade, albeit often in individual subjects. If some subjects, such as the first language, are compulsory for graduation, it is axiomatic that annual progression through the grades is not certain. I shall return to the problem of acceleration and deceleration in the elementary school later. Suffice it to say here that *continuous progress* has not led to the abolition of grade repetition, and there is no evidence that eighth graders have, on the average, become younger over the last twenty years, although certainly the extreme situations of fifteen year olds in fifth grade have been eliminated, either by regulation, as in Quebec, or by custom.

In contrast, England tends to use a combination of streaming and automatic promotion. Pupils are advanced every year on the basis of age, but they may move up or down a stream. Every stream, and, at an extreme, there may be fifteen in a large comprehensive, secondary school, tends to consist of a homogeneous class. Only the top streams take examinations leading to post-secondary education, and there are few opportunities in a senior secondary school for students who are not so continuing. The program of the top stream examination class is quite different from that of the bottom stream class for slow learners.

The combination of English streaming with American grading in some Canadian schools leads to the peculiar situation of the bright student who "fails" enriched eighth grade. That logistical problem does not exist in England, where technically there is no failing anyway--such a pupil just moves down a stream or two.

In general, I would argue against a combination of English streaming and American grading, as the problems of evaluation are symptomatic of deeper programmatic problems. Admittedly, it may be desirable to separate slow learners from other students, at least in some subjects during the junior high years. It may also be desirable to provide advanced and enriched programs in some subjects. But the accompanying problems of program organization and evaluation in such a partial streaming system are less severe than with a full streaming system.

For example, slow learners may be taught basic skills in language and mathematics. They can be marked on the basis of the program they are following (i.e. on a full A to F or 0 - 100% range). A pass does not necessarily denote the minimum standard for entry to high school, but it is not impossible that some students will actually reach that required minimal standard. They can also be tested on a common school examination, in reading, language, written English, and basic mathematics, and on standardized tests.

In the same way, a junior high class might be scheduled separately in mathematics to cover four years in three, or three years in two. The marks should be keyed to the standard necessary on movement into the applicable high school grade. Scheduling an accelerated class in mathematics does not imply that one class must be set apart for all subjects. If there are three mathematics teachers, three mathematics classes can be scheduled simultaneously, so that students can remain in their regular heterogeneous classes for other subjects.

Enrichment, unlike acceleration, generally does not require a separate standard of assessment. Enrichment may be a regular part of the work of heterogeneous classes, or a specific enriched class may be scheduled in the same way as the accelerated class described in the previous paragraph. In both cases, marks should be awarded as part of the regular work of the class. The basic pass mark for competence will not reflect any of the enriched work. On the other hand, a mark of excellence, A or 80% perhaps, will only be achieved by successful completion of some of the enrichment activities. In this way, the student receives a reward, a higher mark, for doing more. If the enriched class, or group, is marked on a different, usually harder, basis, as very often happens, one ends up with a self-defeating and anomalous situation--the conscientious student works harder to obtain a lower mark than a lazier, less ambitious comrade. It is not surprising that high schools sometimes have problems disentangling elementary and junior high school marks.

Grouping

Grouping of students within the class is very common in the elementary grades, much less common in the secondary grades. It generally serves two rather different functions--and those different functions should be reflected in evaluation practices.

Grouping for Co-operation and Enrichment

Grouping is often carried out to permit students to participate more closely in applied or practical work. In fifth grade, pupils may work co-operatively in small groups on different science or social science projects. In twelfth grade, students work in two's or three's to carry out experiments in chemistry. Sometimes, all groups carry out the same or comparable tasks. Alternatively, grouping is used for enrichment--the tasks are varied with the interest and ability of the pupils.

In either case, and the two cases are not clearly separable, the implications for evaluation are similar. Pupils should be marked according to a single absolute standard. This is not so much a problem where all students are doing the same work. It is more problematic if students are doing quite different work. If some students are working on Australian, and others on African, climatic patterns, comparisons will be difficult if the African climate is more complex. The situation is even more difficult when some students are deliberately given not only different work, but much more challenging work. Teachers sometimes argue that a higher standard should be used in marking work by more capable students. The results of that idea are anomalous. The student who writes a fine, but flawed thesis in history, receives 80%, while the less able students who provides a masterly piece of well-written regurgitation gets 95%. Similarly, in physical education, the poorer student who has learned a very simple routine very well, receives an A, while the more capable student, who carried out a very advanced routine imperfectly, receives a B. Both these approaches to marking are unfair. It is one thing to criticize the weaknesses in an exceptionally good piece of work--that is good formative evaluation--but quite another to give it a lower score than an inferior piece of work. Effort is irrelevant, it is the standard of work that counts. There still remains the problem of sharing marks for group performance, a problem to which I return later.

Grouping by Achievement Level for Acceleration

Grouping for enrichment, most often used in music, history, geography, and science should not be confused with grouping by achievement level, which is most often used in elementary schools, particularly in reading and mathematics. In the latter case, students are learning sequential skills. They are at different points in the sequence, and in addition, they are progressing at different rates. Whereas students grouped for enrichment are taking essentially the same course, students grouped by achievement level are taking different courses. The group of students in sixth grade who are reading at the ninth grade level is not taking the same course, by any conceivable stretch of the imagination, as a group in the same class reading at the fourth grade level.

The situation is entirely different from that of the geography class described previously. Students taking the same course should be marked by the same standards. After all, parents and others seeing the mark cannot and should not be expected to take into account that the 99% in music really means that the individual has made fantastic progress considering he is stone-deaf or that the 90% in tenth grade English means he has a vocabulary of 100 words, but has only been speaking English for two weeks. The situation changes entirely if the course the latter individual is taking is *English as a second language for new immigrants*. Bear in mind also that one can have more than one course in a single class.

To return to the reading class, if students are grouped by achievement level, they should normally be marked by achievement level. That is to say, marks should be given according to achievement made within that group working at that particular level. It is, therefore, possible for a student in the fourth grade level group to receive an A and the student in the tenth grade level group to receive a B (surely not an F if the teaching is adequate), always provided that written notice of the level in which the group is working accompanies any record of the mark. This practice is grounds for confusion, although, to use an analogy, no one expects a college teacher of business English to use the same standards as a teacher of honors English in university. Because the system I am advocating is likely to be confusing to some parents, it is vitally important to provide some more absolute measure beside the mark. Thus, accompanying the home report mark of B, in the case of our reading student in sixth grade who works fitfully at the tenth grade level, should be a report showing the approximate level of achievement of the student on a standardized test. Similarly, our high achieving and hard working slow learner who received an A while working two grades below level would also be given an indication of the standardized test level, so that the parents would not be able to claim they were being misled.

Marking in Groups

A quite different and very thorny problem is that of sharing marks for co-operative work. Two good principles conflict--and we would do well to think through that conflict with some care. First of all, a general principle underlying suggestions in this book, is that the *evaluation of student achievement should fairly reflect the individual's actual achievement*. Thus, marks for attitude, effort, and attendance have no place as a part of the achievement mark, essentially because they are frequently unrelated to it (and it does not make good sense statistically to add two generally unrelated components). The marks are more for the product than for the process. That does not imply that marks should be based mainly or entirely on final examinations. After all, the examination is only one way of measuring what people have learned. And I am not suggesting that the components of learning should not be tested and marked. For instance, even though one's goal in French is to have learners understand, speak, read, and write French, it is perfectly reasonable to test vocabulary on the assumption that vocabulary is a necessary condition for those ultimate skills. It would not be reasonable to include, as a part of the mark, attitude to the learning of vocabulary, even though one might hypothesize that a good attitude is desirable for learning French. The fact is that no one who lacks a vocabulary can speak French, but many pupils with unfavorable attitudes, whether recognized or unrecognized by the teacher, do learn to speak French quite well. This is nowhere more obvious than in French immersion situations. Now, of course, few school marks from a teacher are pure measures of achievement, completely free of bias from attitude and effort. The student who turns work in on time gets better marks than the student who forgets. The careful, neat writer tends to get better marks than the equally knowledgeable but untidy one. However, the fact that marks inevitably include, indirectly and unintentionally, a non-achievement component, does not argue for the intentional inclusion of unrelated components that can be excluded. The fact that one tardy student may miss something, and fail to learn, is not an argument for taking marks off the work of another tardy student who has learned. The fact that intelligence contributes to learning, is not an argument for giving marks for intelligence irrespective of whether learning has taken place.

The second corollary of the first principle is that students are marked as individuals. Nearly all marks in school are, in part, competitive. Even if one has a test for graduation in motor mechanics, which most students usually pass, the test loses its significance if no one ever fails. Marks are the basis for the sorting and allocation of young people.

The second principle, and one that conflicts with the first, is that *in many school programs co-operative work is very important*. Thus, highly individualized efforts may sometimes be counter-productive. How then does one provide an individual mark reflecting achievement when a major objective has been co-operative effort? Or should one?

Part of the problem can be resolved by accepting that not everything done in school has to be marked. Extra-curricular activities, for example, are normally not marked, and are, nonetheless, successful. What differentiates extra-curricular activities, which are usually based on co-operation, from in-school programs is their voluntary nature. Thus, voluntary activities in class do not always need to be marked. Students in class can be encouraged to work together on various projects and activities without necessarily being rewarded by a mark. The extrinsic incentive is certainly important but it is not the only incentive. However, in general, most major, compulsory activities in class are, and should be, marked. They should not be compulsory unless they are closely associated with significant objectives. And marks should reflect the achievement of significant objectives.

There remains then the central problem of marking co-operative efforts. Some teachers simply mark the product and give all students responsible equal credit, using the rationale that the results reflect the true meaning of co-operation. The objection to that approach is that it fails to take into account the fact that the marks co-operatively earned are used to show individual differentiation rather than group differentiation. The approach conflicts with the first principle. Once a play has received its applause, nothing happens to the cast. But the individual marks from school do go forward to classify, promote, and certify the student. A fine actress can survive participation in a poorly received play, but a student may not be able to survive the low mark that counts against her.

Very frequently, individual contributions to a joint effort vary greatly. Teachers will sometimes try to reflect that variation by giving different marks for joint presentations depending on the contribution given by different individuals. Thus in the case of a joint discussion, a co-operative gymnastics display, or short play, teachers may give half the weight for the total effort and half for the individual contributions. This method is also open to serious criticism. The gym display is the final outcome. One student may be a star in the display but may have hindered the co-operative effort by failing to train and practise properly with the others. That individual may be the cause of the loss of points to the others, while being a star performer herself. This case illustrates clearly the clash of the two principles--marks should reflect individual achievement, but co-operative activities are more than the sum of individual activities.

There are several ways to tackle the problem, none of which is perfect. One approach is to give the students themselves some power of recommendation of the distribution of marks. Let us suppose the teacher decides that a co-operative gymnastics display by five people is worth, overall, 75% (out of a total possible 100%).

Table XIII

Marking Plan for Gym Project of Five Students

	Precision of Performance	General Appearance	Imagination	Practice	Total
Pupil A					
Pupil B					
Pupil C					
Pupil D					
Pupil E					
Total	200	50	50	75	375

Jointly the students have earned 375 marks (5 X 75). The students are asked to complete Table XIII by distributing the marks given by the teacher for the various criteria. No student may earn more than 100%. A variation would be to have a different group of students carry out the same procedure. The final decision, and responsibility, lie, as always, with the teacher. There will be considerable pedagogical advantage in having students carry out the task if they approach it sensitively and seriously. There are fewer problems with students evaluating others in a collective situation than in one where a group evaluates the performance of an individual. One great advantage of involving the students is that they will come to appreciate the objectives of the exercise more clearly, and that appreciation may make them try harder.

Another variant is for the teacher to divide the marks directly herself, using the same table. This is difficult and it is unlikely that the teacher is able to make reliable estimates of individual contributions in group rehearsals. Students will be alert to anything that smacks of favoritism. The teacher may also decide to allocate the marks, but in consultation with the group.

Yet another approach to the problem is for the teacher to accept the problem and make the most of it. If students of equal, and if possible, complementary ability are placed in groups together, they can be told in advance that the marks will be shared equally. This approach provides a strong incentive for genuine co-operation and therefore has a place in formative evaluation. But, it still permits one individual to make things difficult for others. Whatever approach is used, it is important that students know in advance exactly how they are to be evaluated.

Product and Process

One of the suggestions I have made for the evaluation of co-operative work may appear to contradict some of the points made in Chapter III. If it is illegitimate, the reader may wonder, to mark attitude as part of the achievement mark, why is it legitimate to mark practice for a gym display? If it is acceptable to mark appearance in the display, is it then acceptable to take marks off if students forget to bring their gym clothes to class?

Obviously there are fine lines that have to be drawn. I am certainly not suggesting that teachers should avoid teaching process. Certainly, the correct process is vitally important to the learning of many skills. If one is teaching the crawl in swimming, or problem solving in mathematics, it is absolutely reasonable to mark, in unit tests, the appropriateness of the method used, given that the method is a major objective of the unit. Generally speaking, I would expect there to be a lesser emphasis on method in end-of-term examinations. In the final swimming test, the evaluation is likely to be based on an overall, holistic appreciation of the crawl or the overall ability to solve problems successfully. The higher one is in secondary school the more emphasis there is likely to be on the result. In twelfth grade, spelling is much less important than the overall ability to write a good essay, but spelling may nevertheless be an important component of the fifth grade language mark. The ability to solve problems is more important in twelfth grade chemistry, although competence in setting up and reporting experiments may be more significant in tenth grade.

To turn back to my example in physical education, overall attitude is an inappropriate component of the marks for several reasons: it is not very strongly related to performance; it is difficult to measure; it is easy to fake. Similarly, students may be obliged to wear gym uniforms but that obligation has no more to do with individual achievement than remembering one's books has to do with academic achievement, or buying a ticket to the basketball tournament has to do with skills in basketball. There are other rewards and sanctions that can quite legitimately be employed to make students wear uniforms. Not only that, deduction of marks can easily be counter-productive. A student may decide that he does not mind if he fails physical education and therefore will spend the rest of the year sitting out. When it comes to appearance in the gymnastic display, the clothing and general cleanliness become an integral part of the final product and should be marked, just as one might give a mark for costumes in a class play.

The general principle behind these admittedly fine distinctions is that marks should reflect the achievement of the educational objectives. Factors that are unrelated (statistically or conceptually) should not be added and marks should be used as little as possible as a substitute for discipline. Students should not be given the impression that the only thing required of them is attendance. It may be argued that an electrical company looking for an apprentice electrician is just as interested in attitudes, attendance, punctuality, and reliability as in knowledge about electricity. That may well be so. Let us be honest about it then, and report directly on some of those qualities--but let us not call it electricity. One might not permit a person who fails to attend regularly to continue in a course. One might suspend the student who is dishonest, rude, and unreliable. One could grade students in effort and reliability as well as in electricity and provide all the grades to the employer. But let us not confuse someone who likes music and singing with a good musician, or someone who enjoys languages and loves Quebec with a good linguist.

The distinction I am making then is less one between process and product than among a variety of objectives. Schools should be honest and straightforward about their objectives, and they should be accountable for achieving them. By the same token, students should be held accountable in terms of their responsibilities. If a process, such as the scientific method, is the cardinal objective in science, then obviously achievement of that objective should be an important part of the evaluation of the course. Similarly, presentation of laboratory reports reasonably forms a part of the evaluation, but a much smaller one. However, when one reaches such matters as the teachers' perception of the student's attitude towards science, the number of interventions made by the student in class, and whether or not the student comes to class, the process is so remote from the goal, that an evaluation of it, particularly an unreliable one, is totally inappropriate as a reflection of the goal. An evaluation of achievement of chemistry should reflect achievement in chemistry. In group activities, the line to be drawn is a finer one, but the general principle remains the same. The process should be evaluated to the extent that it is clearly one of the objectives.

Acceleration, Deceleration, and Repetition of Grades

Deceleration

One of the most difficult problems in the elementary school is the question of deceleration. Let me begin with a few premises:

1) The introduction of continuous progress has not reduced the amount of deceleration in Canadian elementary schools;
2) There is little evidence to suggest that children's achievement benefits from greater repetition or deceleration;
3) The usual reason given today for decelerating a pupil is that it is in the best interests of the child, interests that are extremely difficult for the school to determine.

There are contradictions among those three, I believe, factual statements. If continuous progress allows children to progress at their own rate, rather than being held back by the group, why then is progress through the grades now not more rapid and achievement not greater than it used to be? If children do not benefit academically from deceleration, why not let parents decide if children will benefit socially from deceleration?

I suspect children are usually decelerated for two good, but usually, unstated reasons, not entirely dependent on their own academic or social benefit;

a) Children work better with an incentive, and automatic promotion removes that incentive;

b) Automatic promotion provides teachers with a large range of cognitive skills within a single class, particularly in the junior high grades, before course differentiation takes place in high school.

With respect to the first reason, it can be argued that England manages elementary education well enough without grade repetition. However, in England, children are often faced with the far more powerful impact of streaming--more powerful because many a child survives grade repetition to complete high school, while the child pushed into the lowest streams in England rarely lasts beyond the age of compulsory school attendance. It is true that many English elementary schools are now unstreamed, but parents and teachers know well that stratification normally begins in real earnest by the age of eleven.

Rightly or wrongly, deceleration is very common in today's schools. How can it be justified? Most often, schools say they take into account two criteria--academic progress and social, personal considerations. The problem is that frequently, neither is very clearly defined in a written policy. Parents and pupils may well perceive that if one complains loudly enough, one's child is not decelerated. What's wrong with that? Two things: if the policy is that parents make the decision, it should be publicly stated; and, bowing to parental pressure is likely to accentuate social class differentiation in the school, with middle class slow learners progressing on schedule, and working class slow learners being held back. Further, if the real reasons why children are decelerated are more to do with teaching convenience and organization, and incentives for other students, it becomes obviously unfair if the policy is unequally applied.

The above points lead to two conclusions:

1) as the effects of deceleration are, at best, uncertain, the process should be used sparingly;
2) the deceleration policy should be clear.

Sample Deceleration Policy for an Elementary School

General Policy

In general, pupils advance through school according to their chronological age, with program variations being made to accommodate different levels and rates of learning. Pupils are decelerated exceptionally, when their level of learning is such that progress into a higher grade would create significant problems for the pupil and for the school.

Procedures

1) No recommendation for deceleration will be made unless the parents and, normally, the student are informed in writing of the possibility before March 31.
2) Recommendations for deceleration are of two types: a) mandatory, b) permissive
   a) Mandatory recommendations may be appealed by the parents, but only evidence of error in the interpretation of factual evidence will change such a recommendation.
   b) Permissive recommendations are left to the discretion of the parent for final decision making.
3) Only permissive recommendations will be made until first grade, i.e. there will be no mandatory retention in kindergarten.
4) Mandatory deceleration will take place no more than once during elementary school for any child.
5) The following factors will be taken into account in making recommendations for deceleration:
   i) Level of achievement. Except in first grade, no pupil will be recommended for deceleration who is not more than one year below grade level in reading and/or mathematics. In first grade, pupils who have learned basic reading and number skills will not be recommended for mandatory deceleration.
   ii) Learning ability. Students of normal or above-normal ability will not normally be recommended for mandatory deceleration. Only when all other reasonable steps have been taken to remediate the learning problem will such pupils be decelerated.
   iii) Effort. Students who, in the judgement of the teachers, consistently make their best effort will not normally be recommended for mandatory deceleration.

iv) Chronological age. In the primary grades, extreme youth for the grade level will be an additional factor used to support a recommendation for deceleration based on the first three criteria.

v) Socialization. Inability to work and socialize well with fellows of their own age will be an additional factor used to support the recommendation for deceleration based on the first three criteria. It will *not* be used as grounds for exception from the first three criteria.

vi) All assessments made by the school relevant to its recommendations will be available to parents in writing.

The above policy is generally self-explanatory. Achievement is the basic criterion. After all, factors such as effort and socialization are difficult to assess and may vary greatly with the social environment, most notably the teacher. If an average school follows the policy sincerely, and if it has made intelligent efforts to accommodate varying rates and levels of learning, the deceleration rate should be less than 3% per annum. Even then, 20% of pupils may have been decelerated by the end of eighth grade.

The policy is designed to avoid some of the possible abuses of deceleration: i.e. making five year olds repeat kindergarten because they have not learned to play with friends (why will they learn better in the less mature atmosphere of kindergarten?); the punishment of bright students whose reaction to a dull program is rebellion; the holding back of well behaved slow learners in sixth grade, while the recalcitrant ones are sent on to junior high; the use of vague social factors which are difficult to assess reliably; a two category system for the middle and working classes; and the avoidance of hard achievement data.

In practice, most deceleration will probably take place in first grade, where it is very important that children learn the basic skills, and in sixth grade, where it is important that pupils are prepared for the lower level of spoon-feeding they are likely to encounter in junior high school. There will be some deceleration in seventh and eighth grades, particularly of students who are planning academic careers in high school, but who are lacking some of the basics.

Acceleration

It is ironic that the advent of continuous progress (working at one's own rate) has been accompanied by a decline in the practice of acceleration. One of the reasons is undoubtedly the problems that arose when schools tried to introduce continuous progress just at the time that tests of learning ability and of achievement were falling into disrepute. One hears stories, although I have not seen them authenticated in the literature, of large numbers of pupils being accelerated during the late 1960s and early 1970s, only to meet serious problems when they reached high school. Acceleration can easily become a status symbol, and the more common it becomes, the heavier the parental pressure.

If there are at least two reasons for deceleration, there is probably only one valid reason for acceleration--the inability of the school to cope adequately with the student's level and rate of learning without it. Generally speaking, we do not want to rush students through high school and get them to work or college a year earlier.

One can easily see then the problem to the school in acceleration. It must admit that it cannot cope without it. Worse, the school gives implicit support to the idea that the student is "bored"--a favorite word of ambitious parents and lazy children.

In advocating acceleration, I am certainly not supporting the wholesale advancement of pupils who reach a grade level of achievement one year ahead of their chronological age. That leads to six and seven year olds being pressed into rapid achievement in reading and number, only to be advanced into situations requiring abstract thinking and reasoning of which they are quite incapable. The child advanced in first grade must still be able to compete in sixth and tenth grades. It is particularly painful for the child of average ability who is accelerated, only to find, in fifth and sixth grades, that most of his peers have powers of abstract thought well beyond his. From being classed as excellent, he is soon seen as being mediocre, or worse.

A sound acceleration policy will have two elements: no student with less than superior intelligence will be accelerated; in addition to superior intelligence, the accelerated student will normally be about two years ahead of grade level in the basic skills, over one year ahead in mathematics, and at least two in reading. In addition, no pupil will be accelerated against the express wishes of the parents.

That said, acceleration can be extremely beneficial to the bright student. Some bright students tend to atrophy, they seek a dull grey anonymity in the middle of the class, where their brightness will not be remarked upon. The challenge of advancement can be useful. Educators should recognize the limits of even the good teacher and the provision of enrichment. A pupil well in the superior range of intelligence, of I.Q. 130, will, by the age of twelve, typically be scoring above the tenth grade level in tests of reading achievement, and the eighth grade level in mathematics, provided there has been good instruction. That child is usually better off in seventh than sixth grade.

The Use of Examinations

Most secondary schools use examinations once, twice, or even three times a year to help evaluate students in all or most subjects. The loss of instructional time as a result of lengthy and complex examination schedules must be set against the value of encouraging students to study for the examinations, synthesizing the important things they have learned. Some educators complain that examinations merely force students to memorize endless facts, facts that they instantly forget after examination. There is no question that some examinations demand little beyond factual recall, and that we all forget a good proportion of what we have learned. On the other hand, teachers' tests are often factual too--so the abolition of exams only replaces a more scrutinized type of evaluation for one that is scarcely scrutinized at all. My recommendation is that good examinations, once or twice a year, be combined with other means of assessment, approximately on a 50:50 basis.

There are often arguments that the nature of different subjects demands either fewer examinations or that, in some subjects, that examinations count less. One is always suspicious when a department wants to retain examinations for others, but abolish them for itself. With a very occasional exception, virtually all subjects can, and should be examined, including vocational and technical subjects. With the possible exception of art, there would not seem to be strong, convincing arguments for much variation in the proportion of marks to be attributed to examinations. Examinations are not necessarily pen and paper affairs. In physical education, examinations may be in the form of tests of physical fitness. Practical examinations are often imaginatively used in subjects like electricity and automobile mechanics. In the academic subjects, open-book examinations, or open-book sections of examinations, are a good cure for too much memorization.

One strange custom, common in some parts of Canada, is the exemption of students from examination if they have achieved a certain level in their term marks. The argument is that exemption encourages students to work hard on their daily work, in which case a formal examination becomes superfluous. I find that argument singularly unconvincing. The better students are precisely those who need the training of writing examinations, who will profit from being forced to synthesize what they have learned, and who most require fair and equal treatment--they will compete most fiercely for limited positions in university and college programs. A system of exemptions encourages the idea that examinations are a punishment one must undergo, when a good examination gives a successful student a real sense of accomplishment and a feeling of liberation and exhilaration.

Summary

The School Level

This chapter has examined some of the major policy issues that affect the administration of student evaluation in the school. Although the individual teacher cannot decide most of these matters independently, they are of importance to teachers because they are matters that can be determined collegially and professionally.

Mark Allocation

The policies by which marks are allocated are of crucial importance. Although scaling is not a desirable practice at the level of the classroom--it reduces the incentive for overall collective achievement--it does not follow that every teacher should choose an independent and arbitrary path. There should be comparability of standards among teachers and among subjects. Marks should be keyed to specific objectives and levels of competence. Although teachers' tests should be designed so that they parallel the importance of the various objectives, they should be servants rather than masters. If tests are imperfect, creating unanticipated difficulties for students, it is not unreasonable to determine a new, lower pass mark based on the reality of the skills actually tested, rather than on one's original intentions. However, such mark "fudging" should be distinguished from scaling marks up just because "too many failed." If too many failed to achieve the objectives, the objectives may be inappropriate, the students may be taking the wrong course, the students may not have worked hard enough, or they may have been badly taught. In any event, nothing is gained by certifying the incompetent as competent. Mark adjustment is normally only justifiable when the test fails to assess the achievement of the objectives very accurately.

Groups and Streams

There is no reason why different levels and different streams should be marked by the same criteria--in fact, every reason why students should be marked against criteria appropriate for the course they are taking. However, if the marks are to have meaning, the course must be clearly described.

Group Work

Group work poses a particular problem for teachers because marks are individual and competitive, whereas group work may have co-operative achievement as an objective. A compromise must be found between marking the group as a whole, thereby penalizing the hard-working and co-operative individual within a difficult group, and marking the individual within the group, and thus destroying the incentive for genuine co-operation. Several suggestions have been made to achieve a compromise, one being to create homogeneous groups, another being to score the components of co-operation as well as the final product, and the third being to have students take responsibility for sharing fairly a set pool of marks.

Deceleration and Acceleration

Deceleration fulfils a useful function in the school--it makes organization a little easier for the teacher and provides some incentive for students to work harder. A carefully worded policy must be worked out that not only can be fairly applied but is seen to be fair by parents and students. Lack of achievement should be the central criterion for deceleration. In the case of acceleration, however, rapid achievement is a necessary but insufficient justification. Only students who are both high achievers and of superior intelligence should be accelerated. They must be able to cope with peers in the future as well as in the present.

5 THE SCHOOL AND THE PUBLIC

My purpose in writing this book has been to attempt to deal with some of the commonest, most difficult problems of student evaluation. Many of the reasons for student evaluation are external to the school. Students are certified as having certified competencies. They are provided the means of access to employment or further education, and they are evaluated in such a manner as to give their parents an understanding of their progress and difficulties.

It must be apparent than that it is insufficient for a school to develop a perfect set of evaluation procedures if it is misunderstood by other schools, by universities, and by parents. It was fashionable in the early 1970s to evaluate young children in terms of "their own rate." This procedure was doomed to ridicule and failure because, obviously, by definition, all children progress at "their own rate," and therefore should all deserve the same mark. Even if one translates the euphemistic "their own rate" to mean the rate "the teacher thinks the child should progress at," any grade given is obviously dependent on the teacher's arbitrary assessment of the best rate possible. In fact, reports of this kind rarely provided the basic assumptions on which they should have been based. Small wonder then, that the public became upset when students with fine report cards in terms of "their own rate," turned out to be very low achievers. It was a small step then for parents to demand simple report cards at the elementary level that would place everyone in a given grade on the same basis--an A is an A is an A.

I have suggested that it is more appropriate to report progress in terms of the level of the group in which the pupil is working. It makes little sense to me to give an 'E' or an 'F' to a hard working student who is making good progress in fifth grade level reading, just because that student happens to have been placed in sixth grade. But, if my plan is to be accepted, it must be clearly explained to all parents. The level at which the group is working must be clearly recorded on the home report and in all school records.

Similar problems arise at the high school level. There, general and regular classes are marked on different achievement bases, and rightly so. Parents do not complain--after all, they usually know what level of course is being taken. However, few schools publish clearly the criteria for movement from general to regular classes, or the reverse movement. Worse still, a tenth grade student of regular mathematics is likely to receive little beyond the mark of, say, 76%. No indication is given of whether 76% is good or bad. Does 76% reflect excellence, competence? Will a continued achievement of 76% provide entrance to university, to college, to a university course in mathematics? Additional help is provided sometimes in the form of a median mark. To be really informative, two median marks should be provided--the median mark of the class and the median mark of all classes following the same course. If the former is 55% and the latter 77%, then an informed parent has grounds for intelligent questions. Does the class in question have a particularly poor group of students, or a poor teacher, or has it been given more difficult assignments? Another way of providing essentially the same information is to give the median mark of the class together with the quintile rating of the student in comparison with all students following the same course. The quintile rating shows whether the student is in the first, second, third, fourth, or last fifth of all students.

There must be written interpretation of marks on report cards. I am not referring to additional anecdotal comments so much as to clear printed explanations. Anecdotal comments do serve a purpose, particularly in the earlier grades. Parents should be told the reasons why, in the judgement of teachers, pupils are failing to make good progress in some areas. Suggestions of steps parents and pupils can take to improve may be valuable. However, anecdotal reports cannot possible substitute for marks or letter grades.

The choice between letter grades and percentages is one that often exercises teachers, and sometimes parents. Both fulfil essentially the same purposes. A criticism often made of percentages is that they give a misleading air of accuracy--how can one possibly distinguish between 59% and 60%? I find that argument a little spurious, although I do agree there can be no real distinction between 59% and 60%. But with letter grades, there still must be an equally fine line drawn between A and B and C and D. With both percentages and letter grades, two consecutively ranked students are graded differently. In fact, percentages have an advantage in this respect. After all, the parent and pupil can easily see that little change is signified by a change of mark from 60% to 59%. The move from C to D, or D to F appears wrongly to be much more definitive.

The advantage of letter grades is that they free one from the tyranny of marks. It is incongruous and contradictory, but not, unfortunately, implausible to imagine a teacher saying, "I'm really sorry about Joe. He's done good work this term, and I really think he is a passing student. His work is quite acceptable but, I have to fail him as he only made 48%." As suggested in the last chapter, failure should be based on failure to achieve the minimum standard. It should not be dictated by the chance of a difficult task.

That problem is more easily soluble if one is using letter grades. It is more difficult to average letter grades from several subjects, but average marks have comparatively little general utility. They are necessary to obtain college and university entrance in some provinces. It is also useful to award those who achieve best academically with prizes. For these purposes, with the agreement of the universities and colleges concerned, letter grades can become grade-point equivalents. One can even turn grade-points back into numbers that look like percentages, if universities cannot handle grade-points. These matters are more of interest at the level of the school district or province. It would be difficult for a single school to change alone.

Conclusion

Student evaluation is intended to provide evidence of the progress of learners towards sets of explicit or implicit objectives. That evidence is useful to the teacher to assess progress, to the pupil as an incentive, and for purposes of classification and certification, to the parent, to the employer, and to educational institutions. It is a means of communication. As such, it should be clear and unambiguous, and there must be permanent, written evidence of that clarity. It is my belief that learners will find school more worthwhile, parents will be more supportive, and teachers more rewarded, if greater emphasis is placed on the clear and fair evaluation of unambiguously stated objectives. It is my hope that this book will help bring about that desirable state of affairs.

INDEX

acceleration, 119, 126, 130, 133
advanced courses 111-112
aesthetic domain 82
art 89-91
assessment, for instruction 3, 9, 11
attitudes 83-86
automatic promotion 116

bahavior (discipline) 87-89

certifying learners 13
classifying progress 11
co-operation 118-124
continuous progress 116
criterion-referenced tests 31-32

deceleration 126, 133
deciles 28
diagnostic evaluation 22-23
domain-referenced tests 37

emotional domain 82
english 51-53, 72-79, 91-93
enrichment 114, 118
essay questions 72-79
examinations 42-43, 131-132
extra-curricular activities 122
extrinsic rewards 9

formative evaluation 22

general courses 111, 115
geography 50
goals 31
grade promotion 126-131
grouping 118
groups, marking in 120-124

health 68
holistic scoring of essays 72-76

incentives 9
intelligence tests 23-28
I.Q. 23-28
item banks 33-37

learning ability tests 23-28
levels of courses 111
listening 91-93

mathematics 96
moral domain 82
multiple-choice questions 56-71
music 89-91

norm-referenced tests 28-31

objectives, higher level 66
objectives, type of 42-43
oral presentations 51-53, 91-93

paragraph questions 72-82
percentiles 28
physical education 94-96, 122-123
projects 96-98, 123

questions--matching 51, 55
questions--multiple choice 48, 56, 59, 66, 69
questions--oral 51
quintiles 28

reliability 17-22
repeating grades 126-131

scaling of marks 100-102
science 50,68
self-respect 88
social behavior 87
social domain 82
speaking 91-93
spelling 50
spot tests 10
standardized tests 28-31, 32,
streams 111, 113, 116, 133
summative evaluation 22, 33

tests--Cloze 53
tests--criterion-referenced 31-32
tests--domain-referenced 37
tests--intelligence 23-28
tests--multiple choice 48, 56, 59
tests--norm-referenced 29
tests--standardized 28-31, 32
tests--teacher made 33
tests--I.Q. 23

validity 17-22
values 86-87
variations of marks 102-111
vocational education--
  electricity 125

writing 72-79